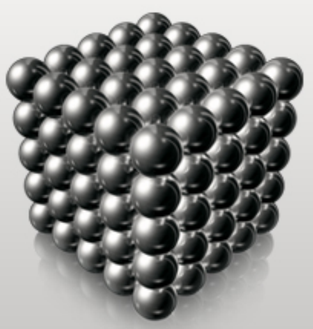
Large-Scale Xserve RAID Deployment

Session 638

These are confidential sessions—please refrain from filming or taking pictures.



Enterprise IT

Large-Scale Xserve RAID Deployment

Introduction

Alex Grossman
Senior Director, Storage and Servers
Apple

http://developer.apple.com/wwdc2005/

Large-Scale Xserve RAID Deployment

Harry Aine
President
SAN Solutions

http://developer.apple.com/wwdc2005/

What We'll Cover

- SAN Solutions background
- SAN Planning and deployment process
- Large scale deployment vs. small
- Small reference design and test results
- Scaling the small design to a large design
- Large scale reference implementation
- Conclusions

SAN Solutions, Inc. Background

- Founded in 1998
- Prior life at MountainGate which was acquired by ADIC
- Perform large SAN/NAS integration
- Specialize in rich media
- High performance, highly available infrastructures
- Customer portfolio includes Sony Pictures Entertainment, Microsoft, Johnson Space Flight Center, Skywalker Sound, NBC, Warner Brothers

SAN Solutions Clients

- Most clients
  - Are more than 10 seats
  - Have storage needs more than 10 Terabytes
  - Are running real-time applications
  - Have both Apple PowerMac workstations and Xserve Servers
  - Highly dependent upon uptime is an understatement!

SAN Planning Process

- Successful large scale deployments must have each of the following:
  - Requirements assessment
  - Storage/SAN design
  - Implementation
  - Documentation
  - Training and support

SAN Planning Process

http://developer.apple.com/wwdc2005/

Customer Requirements Assessment

- What are the business and key personnel goals?
  - For Example
    - Our biggest problem today is we cannot find anything
    - It takes too long to transfer files
    - People within our organization have expectations we cannot deliver today
    - We have too many failures today, too costly, too demoralizing
- What is the risk adversity level?
  - Are they an early adopter of new technologies?
  - Or—are they more comfortable adopting a proven deployment?

Customer Requirements Assessment

- Implementation Scope
  - Workgroup, department, multiple departments, an entire facility
  - Today and in the future
- Customer IT resources
  - What IT staff is available?
  - What is their level of expertise?
  - How is this expected to change over time?

Customer Requirements Assessment

- Workflow
  - How are people doing their jobs today?
  - What expectations do they have with a new system?
  - How readily are the users going to accept change?
  - How temporal is the stored data?
- Application technical requirements
  - What applications will be running?
  - What problems currently exist?
  - What real-time requirements?
  - 7/24/365 application?
  - Compliance requirements?

Customer Requirements Assessment

- On-going support requirements
  - Any downtime allowed
  - Preventive maintenance and SLA monitoring
  - Break-fix time requirements : 2hr, 4hr, next day
  - Help desk requirements
- Data migration requirements
- Implementation timeline and any phasing
- Budgets

SAN and Storage Design

- Select products appropriate to assessment findings
- Base design from reference implementations where ever possible
- Customize as necessary
- Document physical and logical design
- Design should accommodate future deployment phases
- Design should include any monitoring products required to meet support and SLA requirements

Implementation

- Detailed implementation plan includes:
  - Responsibilities and required resources of integrator, customer, and any other third party involved
  - List of deliverables of integrator and any other third party
    - Pilot testing and any other product qualification is common
  - Project schedule and completion criteria
    - Any maximum load tests and monitoring should be detailed
  - Any change control process

Documentation

- Documentation needs to be sufficient to support site
- Customer may have specific requirements
  - Special format
  - Power requirement
  - Network requirements
  - Logical and physical documents
- Customers commonly ask for tested procedure manuals
  - Shut down procedures
  - Cluster fail-over procedures

Training and Support

- Typically customized plans specific to customer requirements
- Training at various levels
  - Fibre channel
  - Storage management
  - Operational management
  - Etc.
- Support plans are tailored to customer requirements
  - Usually have break fix support
  - On-going training for new employees
  - Infrastructure evolution changes

Large Scale Deployment vs. Small

http://developer.apple.com/wwdc2005/

Large Scale Deployments vs. Small

- More a function of infrastructure capabilities
  - Not directly a function of number of servers
  - Or, total capacity
- More directed to multiple classes of data and complex workflows
  - Real-time video, audio, film, other
  - Databases
  - Shared media vs. single user volumes
- Redundancy for high availability
- Homogeneous vs. heterogeneous clients

Large Scale Deployments vs. Small

- For large scale, start small and then scale
  - Tune storage for applications in small environment
  - Then, make adjustments for concurrent access
  - Add clients, scale small environment to large
    - Avoid mixing data classes, use storage affinities
    - Test at full loads over a long period
- What to watch out for with shared files…
  - Number of log in…
    - Storage arrays usually limit number of host connections
    - Memory requirements on client as a function of targets
    - Too many storage connections can increase latency
  - Again, fan out will effect latency at the disk subsystem

Large Scale Deployments vs. Small

- Managing queue depth at client
  - Fairness please, no hogging
- Fabric design and isolation
  - Physical layer problems can saturate fabric
  - Uptime consideration
  - Firmware uploads
- Dynamic vs. static storage assignment
  - Adding and removing storage can require reboots
    - Pre-allocate and mask vs. adding on the fly
    - Some sites never change storage provisioning
    - Some site change on hourly basis

Large Scale Deployments vs. Small

- Backup and recovery, archiving
  - Continuous replication
  - Disk-to-disk
  - Snapshots
  - Users to manage their own recovery
  - Host client agents for cache coherency
- Management tools
  - Pro-active vs. re-active fault isolation tools
  - Must have a effective means to eliminate finger pointing
- Next, we will go through a small reference design and test results to see how well Xserve RAIDs scale

Small Scale Reference Design and Test Results

http://developer.apple.com/wwdc2005/

Building a Small Reference Design

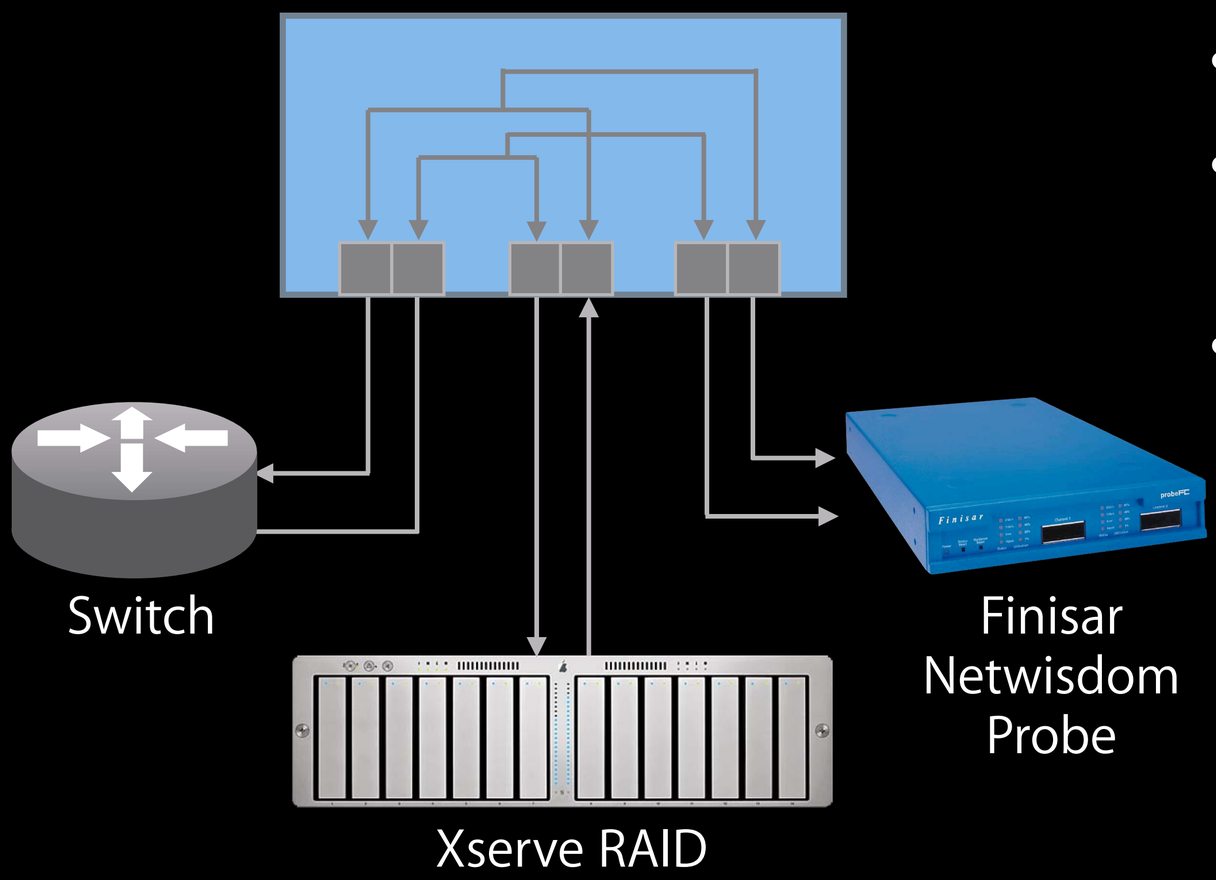
Tools—Doremi Labs HD-SDI and SDI Test Generator

- Multi-format output for HD and SD Video
- Dual-link 2K video output
- Still and moving images embedded time code
- Audio tone and closed caption
- Great video source for rich media SAN testing
- $3,000

Building a Small Reference Design

TAPS—Finisar Netwisdom Fibre Channel Tap

- Passive optical splitter (50/50)
- Allows protocol analyzer or probes to be added without interruption
- New low cost



Building a Small Reference Design

Tools—Finisar Netwisdom Probes

- Collects traffic across fibre channel connection

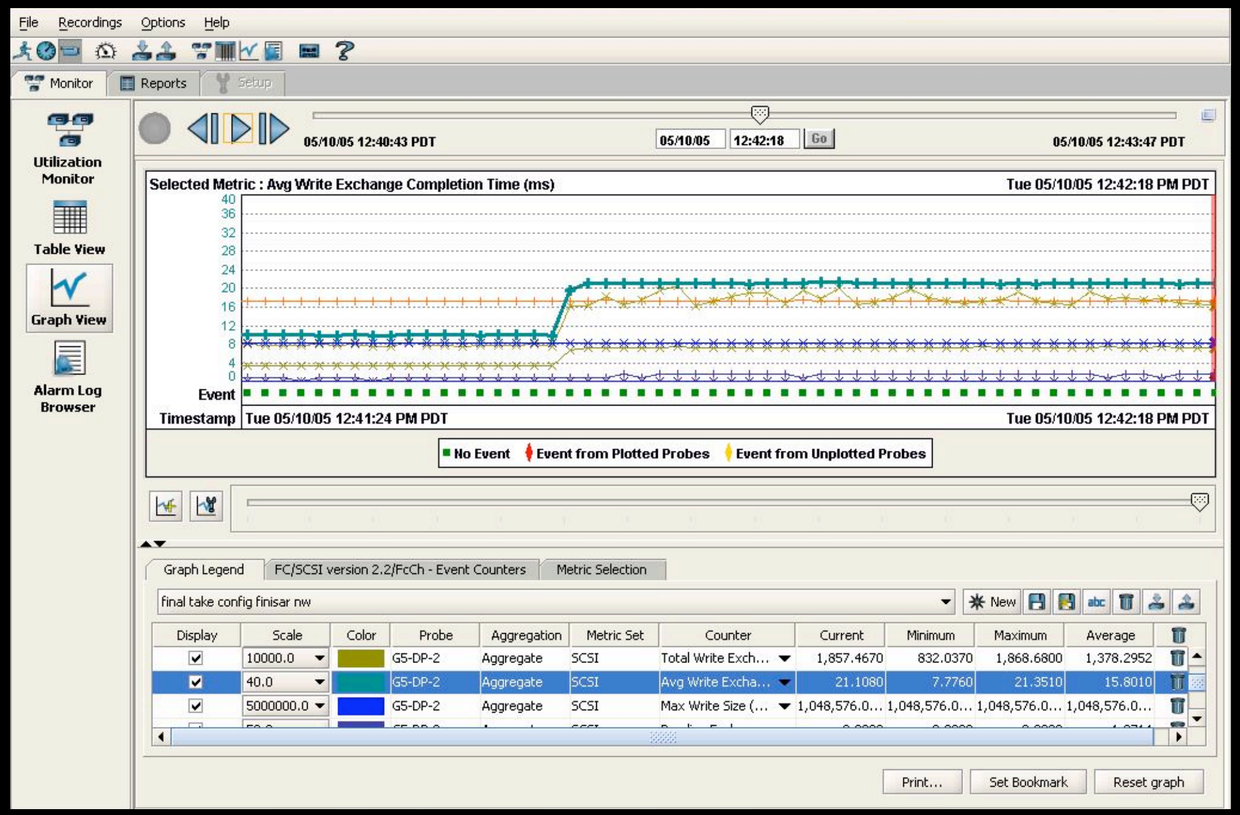
Building a Small Reference Design

Tools—Finisar Netwisdom Portal

- Management Server for Probe
  - Consolidates data from multiple probes
  - Allows remote notification based on policy
  - Report Generation

Building a Small Reference Design

Tools—Finisar Netwisdom Portal



- Views—management application for Probe
  - Shows current and historical traffic
  - Shows physical layer and upper layer protocol events
  - Gives ability to correlate application events to effects on network latency and bandwidth
  - Finisar reports document reference design configurations and loads
  - Current traffic runs give real-time feedback how Xsan configuration changes effect performance

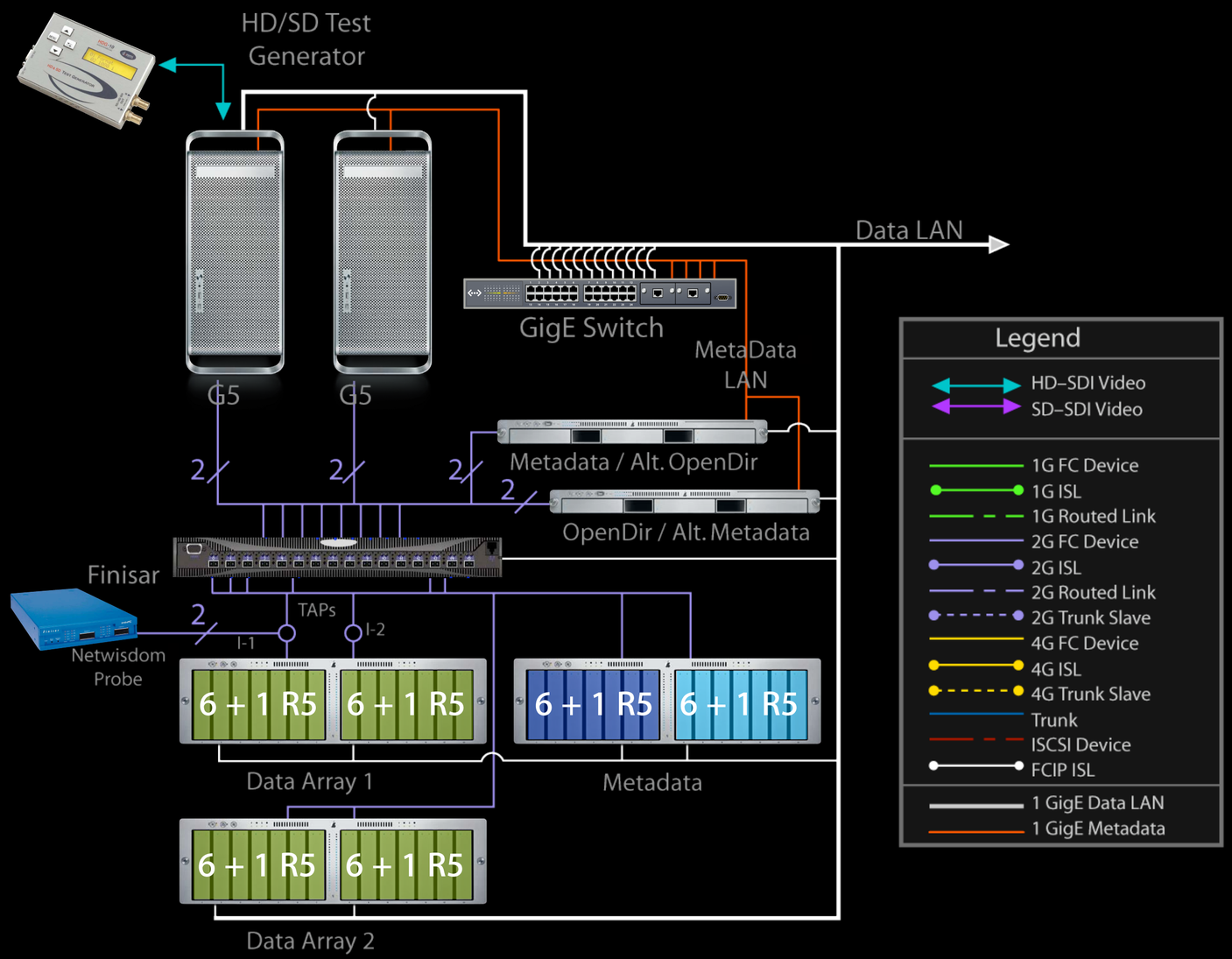
Building a Small Reference Design

Servers, workstations and switches

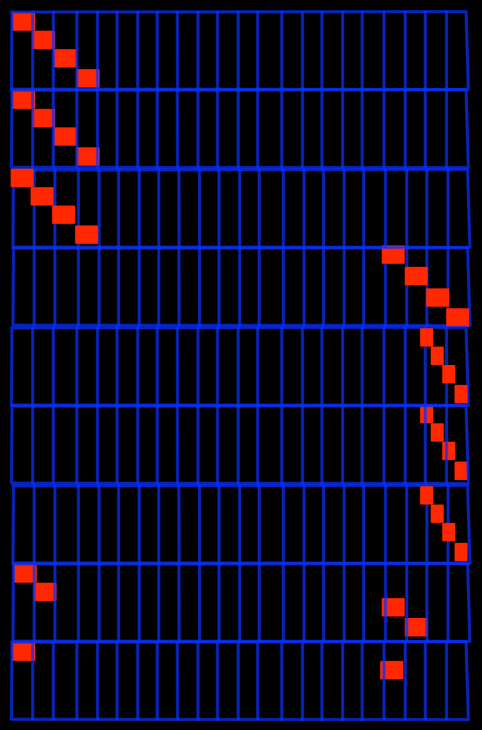
- Brocade 3800 2Gb FC Switch
- Cisco Gigabit Ethernet Switch
- 2x PowerMac G5s
- 2x Xserve G5s
- 3x Xserve RAIDs
- Xsan File System

Building a Small Reference Design

Complete system

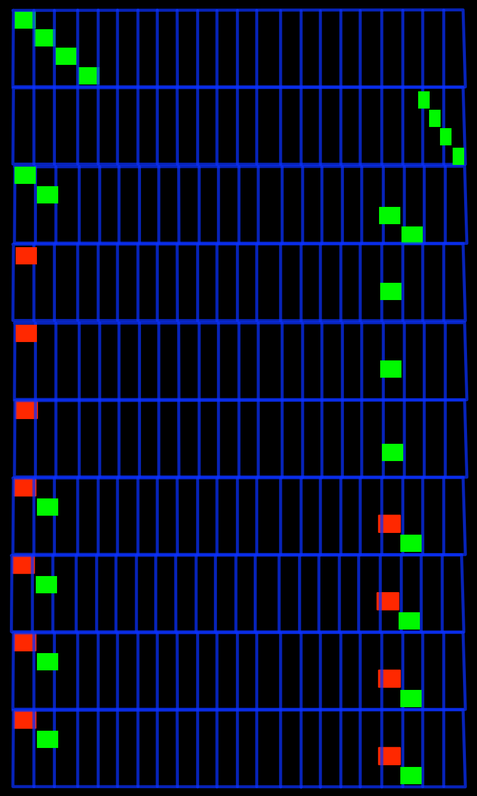


Read Test—Raw IO on 1/2 Xserve RAID

Test Pattern		Pre-Fetch	Total MB/s	Total Read Exchange ms	Max Read Exchange ms	Max Read Size (bytes)	Pending Exchanges
	**1**	64KB	110	3950	50	1048KB	4
	2	512KB	165	3950	48	1048KB	4
	3	8MB	191	3940	25	1048KB	4
	4	8MB	110	3000	48	1048KB	4
	4A	8MB	110	4000	48	1048KB	4
	5	512KB	110	4000	48	1048KB	4
	6	64KB	60	4150	90	1048KB	4
	8	64KB	44	4000	180	1048KB	4
	8A	8MB	135	1900	25	1048KB	2

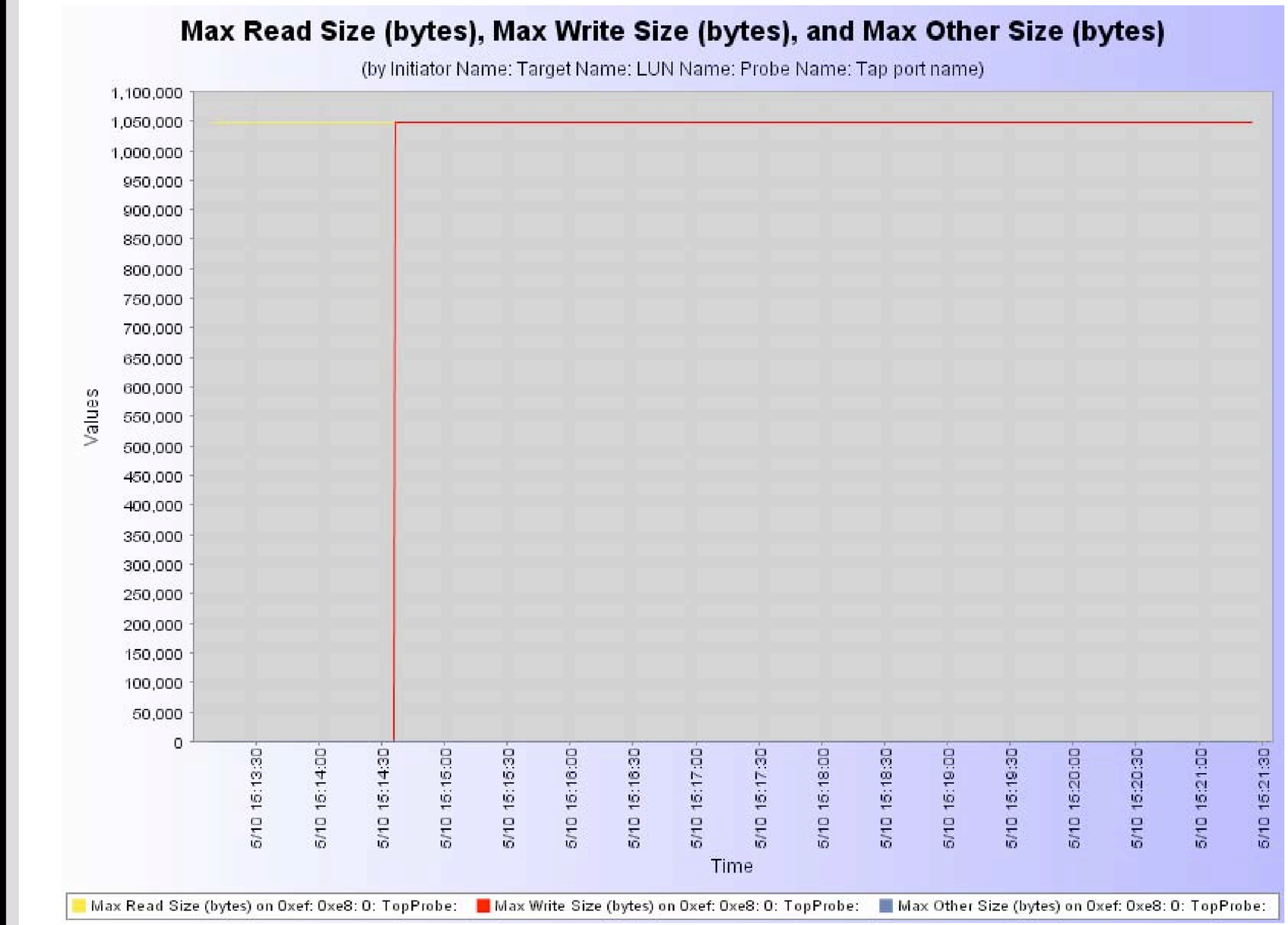
Double numbers for full Xserve RAID Array

Writes and Mix R&W on 1/2 Xserve RAID

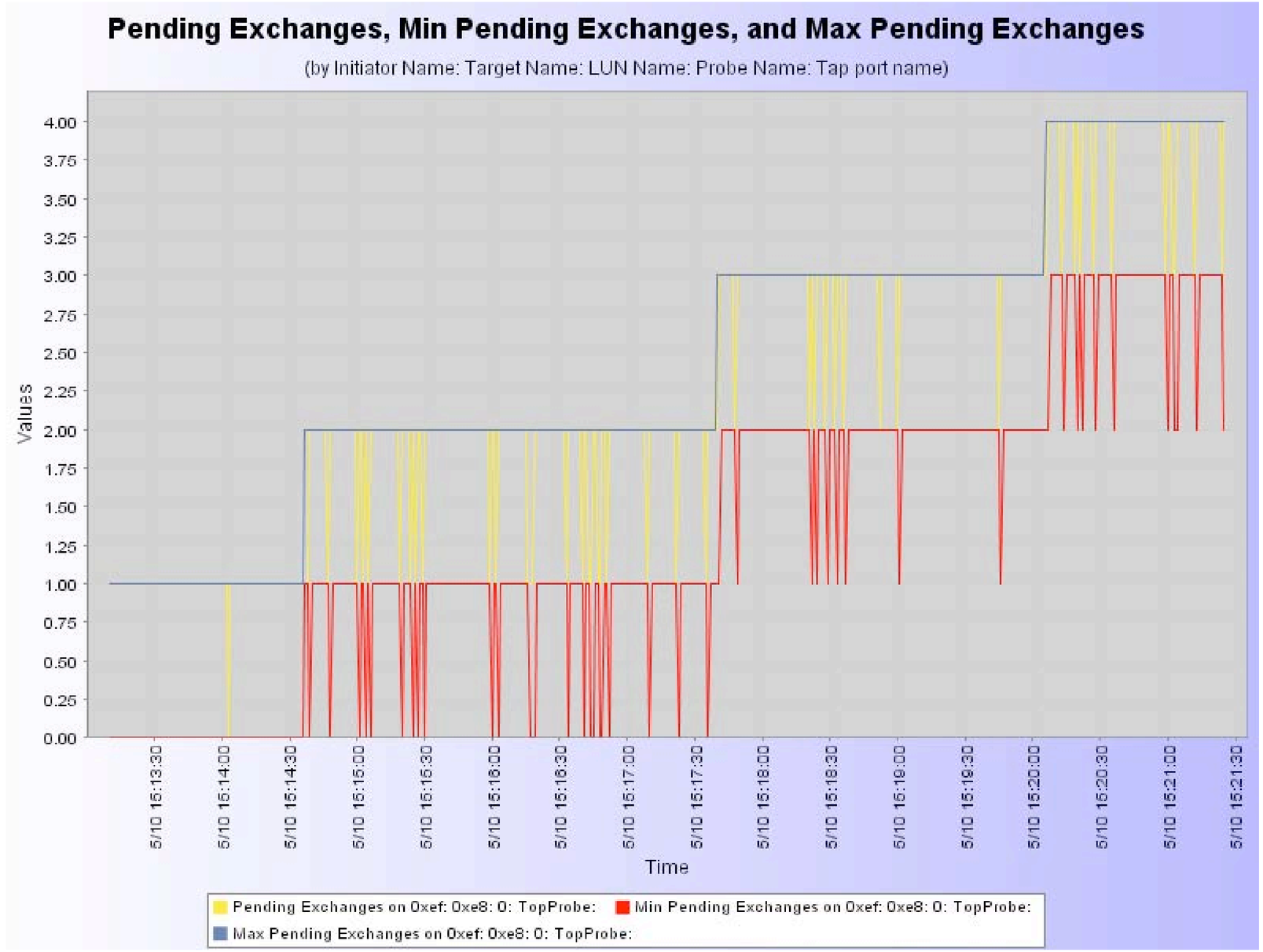


Test Pattern	Pre-Fetch	Total MB/s	Write MB/s	Read MB/s	Total W Exch ms	Total R Exch ms	Max W Exch ms	Max R Exch ms	Max Pend Exch	WCE
10	64KB	87	87	0	3800		50	0	4	Y
11	512KB	87	87	0	3800		52	0	4	Y
12	8MB	87	87	0	3850		52	0	4	Y
15	64KB	87	85	2	900	1100	25	700	2	Y
15A	512KB	95	82	13	900	1000	25	275	2	Y
15B	8MB	115	75	40	950	900	50	50	2	Y
16	8MB	125	35	90	2000	2000	115	50	4	Y
16A	512KB	115	45	70	1800	1900	75	75	4	Y
16B	64KB	110	70	40	1800	1950	75	275	4	Y
16C	64KB	45	35	10	2000	2000	275	100	4	N

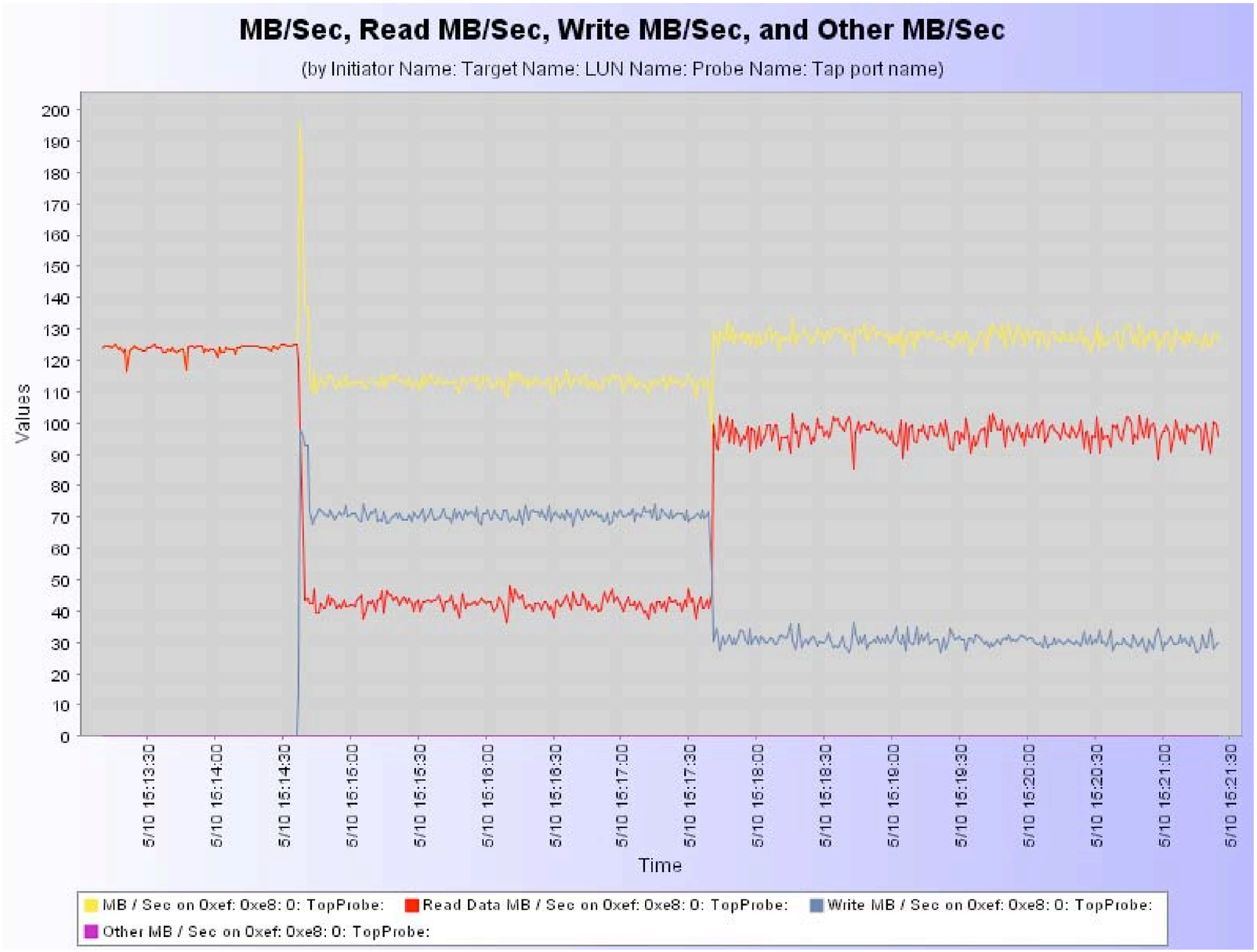
Test 16—Finisar Reports



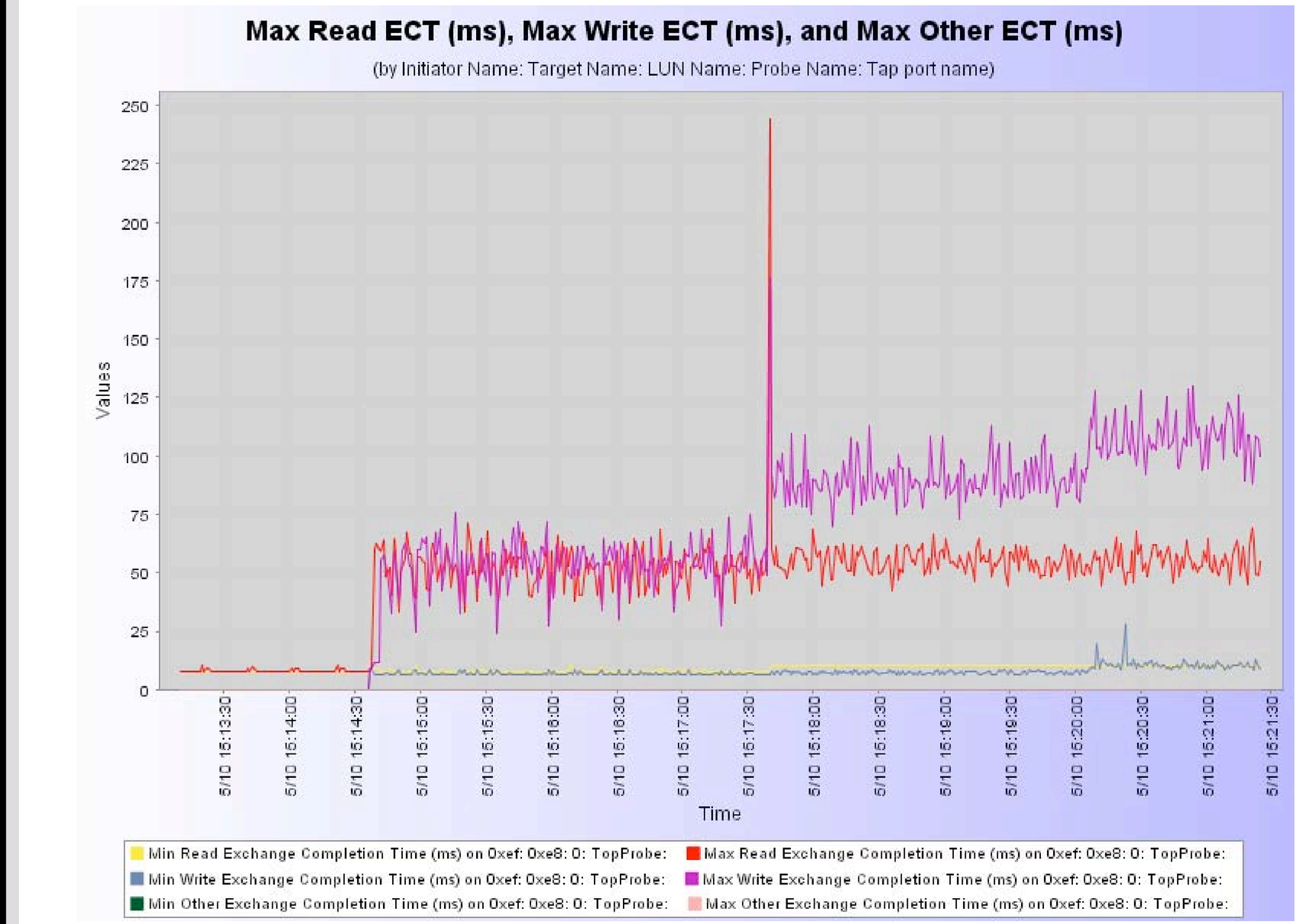
8 MB Pre-Fetch



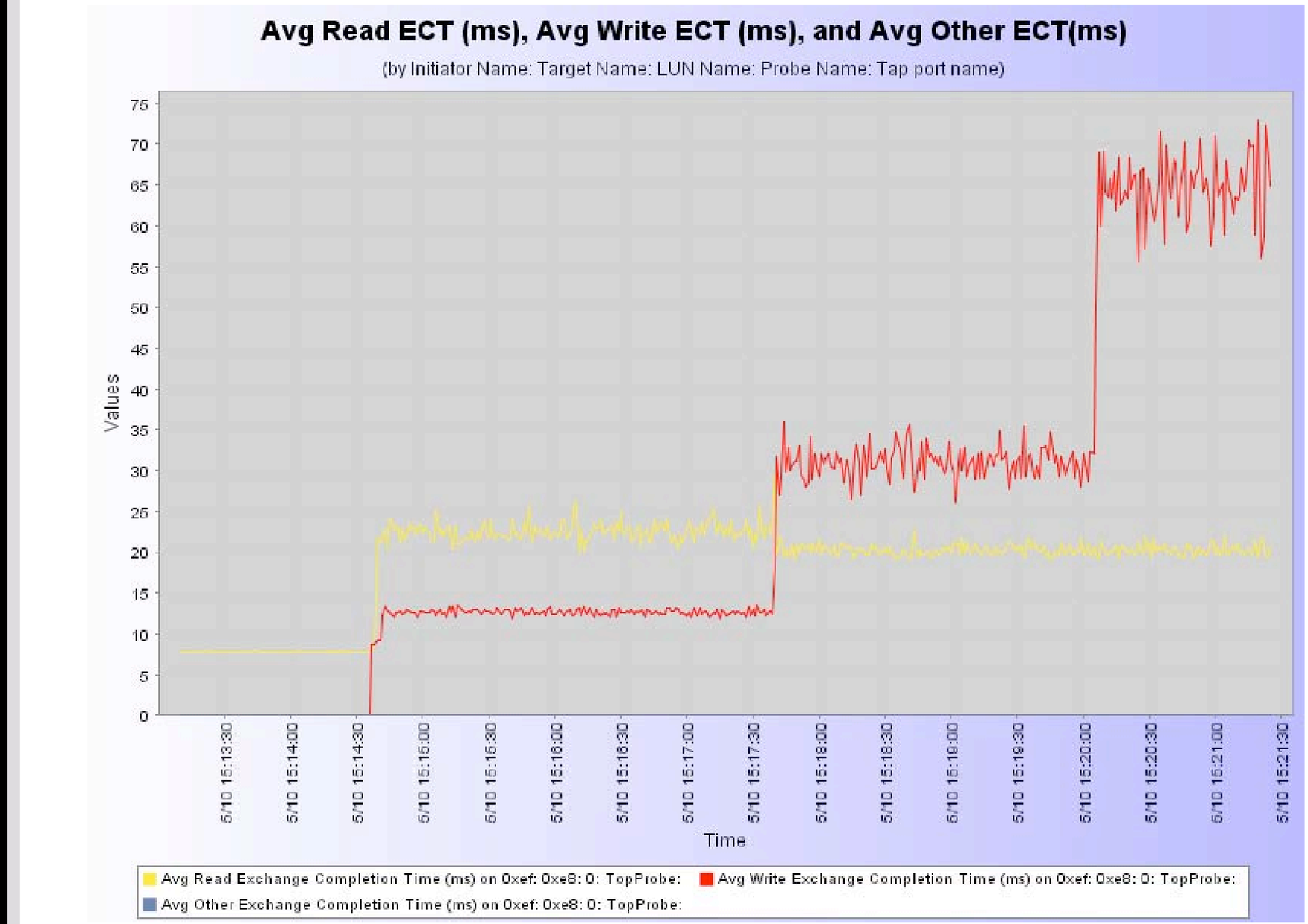
8 MB Pre-Fetch

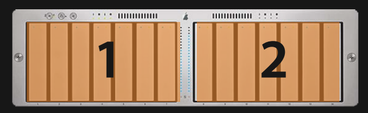


8 MB Pre-Fetch



8 MB Pre-Fetch





Data Array 1

Data Array 2

Block Allocation = 256KB

Stripe Breath = 4

	Target	1	2	3	4
I/O Sequence	1	256KB			
	2	256KB			
	3	256KB			
	4	256KB			
	5		256KB		
	6		256KB		
	7		256KB		
	8		256KB		
	9			256KB	
	10			256KB	
	11			256KB	
	12			256KB	
	13				256KB
	14				256KB
	15				256KB
	16				256KB
	17	256KB			
	18	256KB			
	19	256KB			
	20	256KB			

Test Results—Other Findings

Xsan file system Block Allocation and Stripe Breath

- Determined by Finisar Trace
- Block allocation equals size of transfer
- Stripe breath determines number of I/Os before advancing to next target

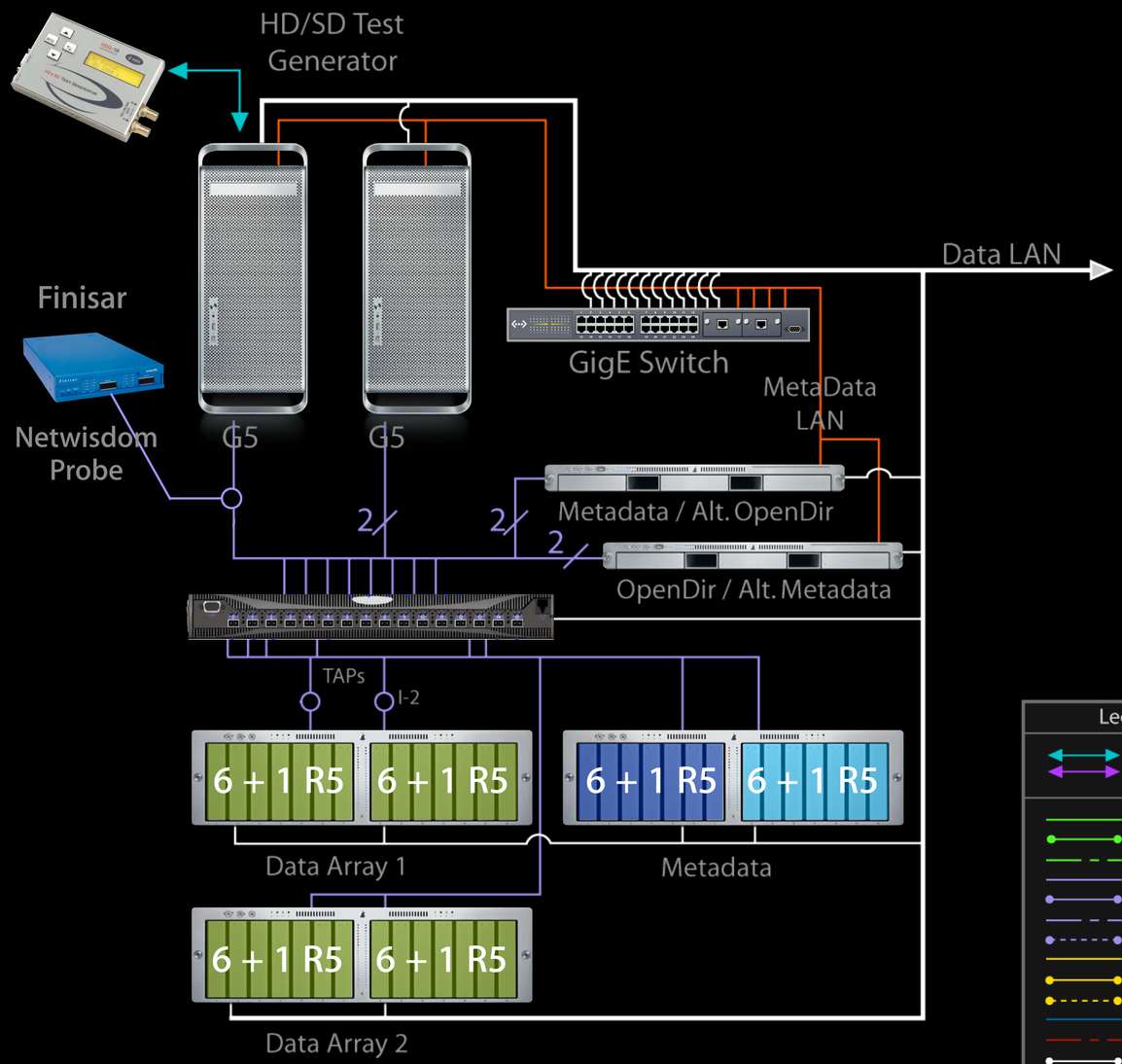
Next, Real HDTV Data

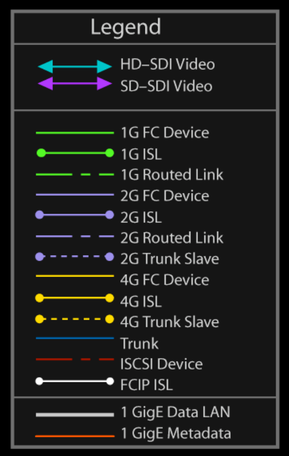
http://developer.apple.com/wwdc2005/

Small Reference Design

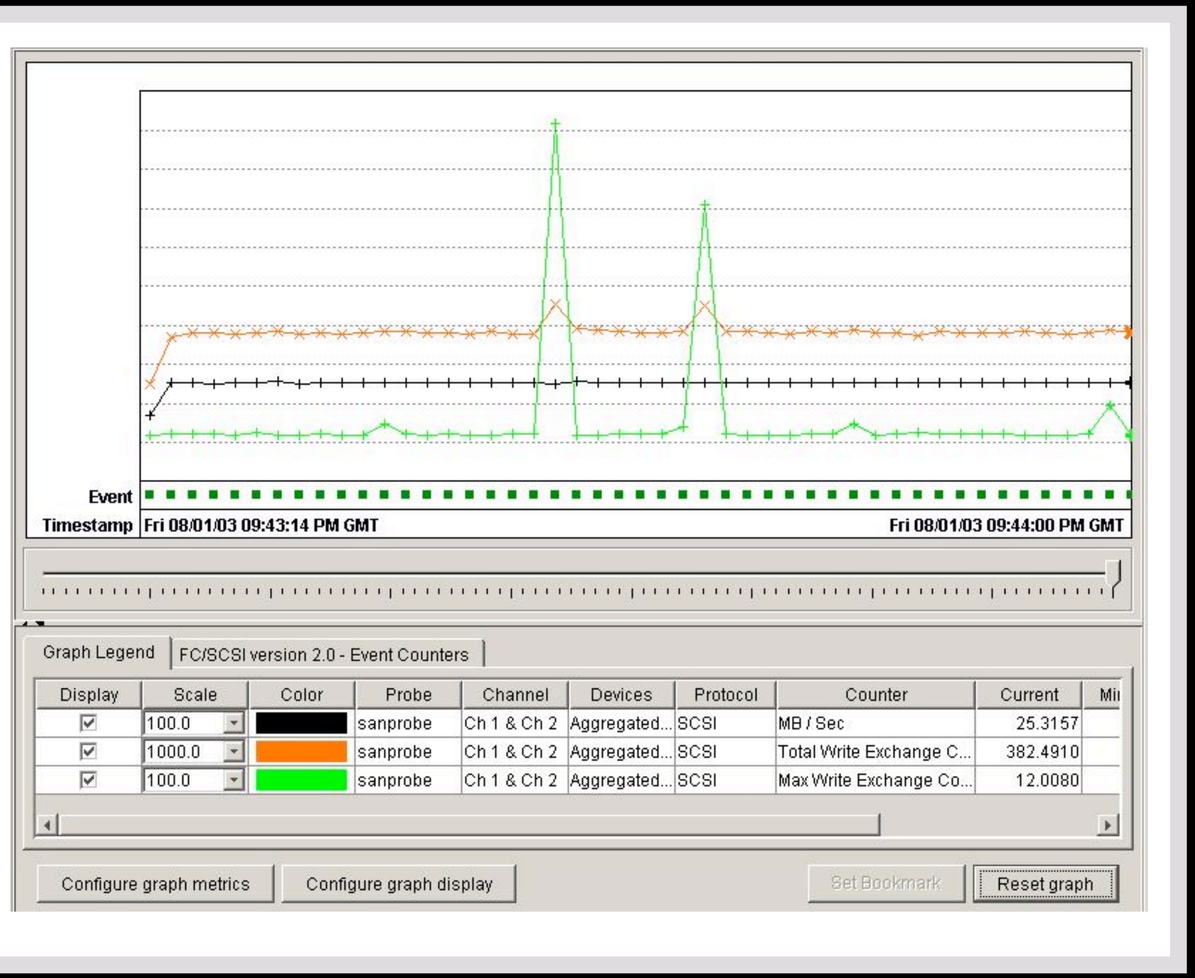
HD Capture Trace

- Move Probe up to G5
- Capture uncompressed HD thru one FC HBA Channel
- Xsan File System across 4 Xserver RAID channels
- Metadata on separate drive
- Review results…





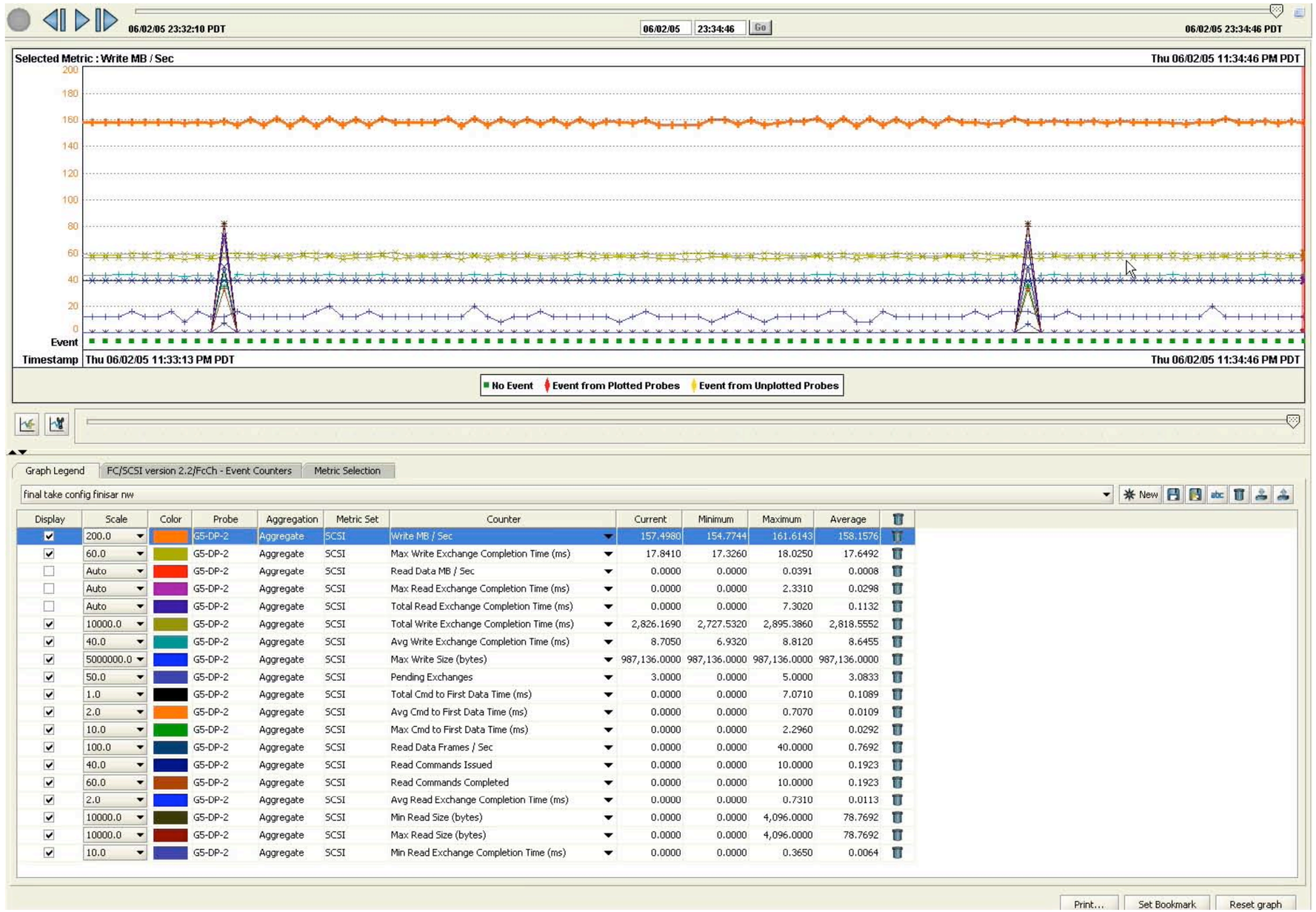
But First, What to Look for…



Typical Array

- SD Trace
- Green line is latency
- Peaks in latency will cause records to drop off
- Leads to unpredictable behavior

HD Capture



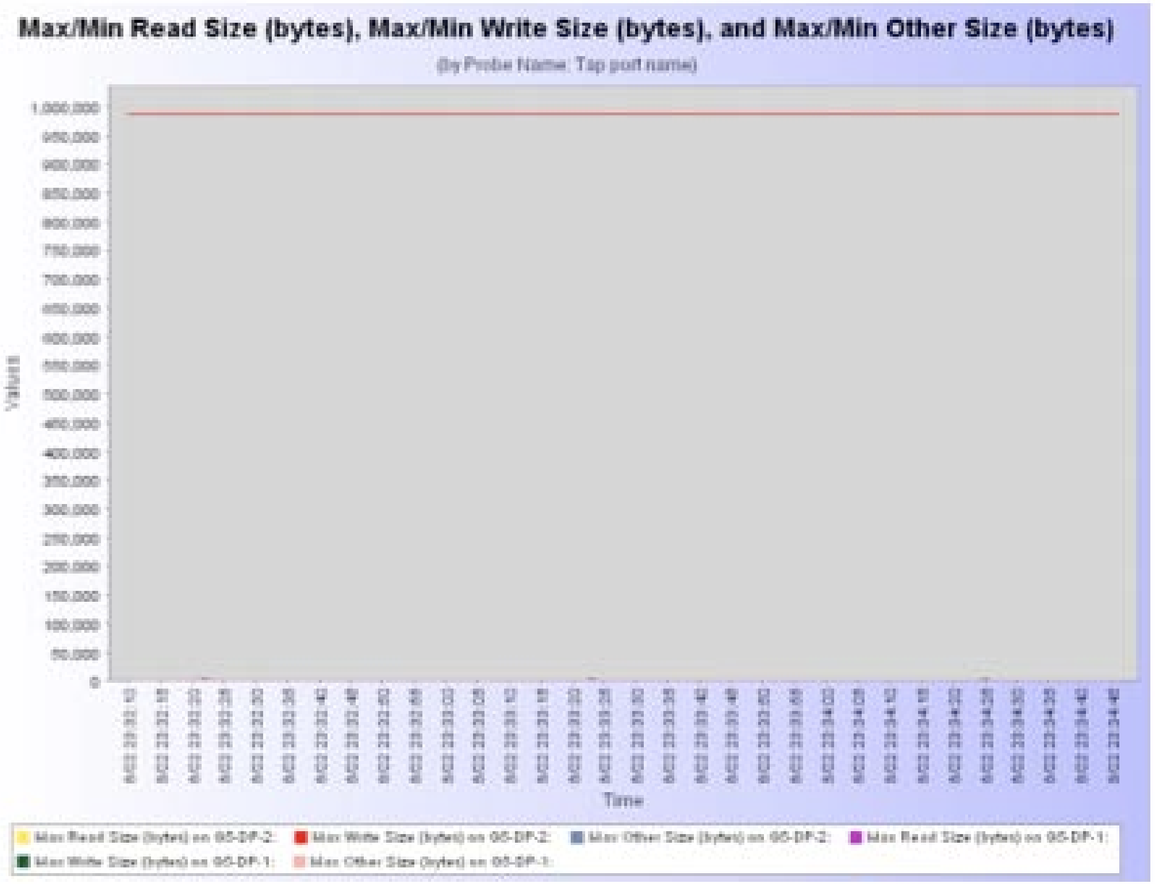
Display	Scale	Color	Probe	Aggregation	Metric Set	Counter	Current	Minimum	Maximum	Average
☑	200.0		G5-DP-2	Aggregate	SCSI	Write MB / Sec	157.4980	154.7744	161.6143	158.1576
☑	60.0		G5-DP-2	Aggregate	SCSI	Max Write Exchange Completion Time (ms)	17.8410	17.3260	18.0250	17.6492
☐	Auto		G5-DP-2	Aggregate	SCSI	Read Data MB / Sec	0.0000	0.0000	0.0391	0.0008
☐	Auto		G5-DP-2	Aggregate	SCSI	Max Read Exchange Completion Time (ms)	0.0000	0.0000	2.3310	0.0298
☐	Auto		G5-DP-2	Aggregate	SCSI	Total Read Exchange Completion Time (ms)	0.0000	0.0000	7.3020	0.1132
☑	10000.0		G5-DP-2	Aggregate	SCSI	Total Write Exchange Completion Time (ms)	2,826.1690	2,727.5320	2,895.3860	2,818.5552
☑	40.0		G5-DP-2	Aggregate	SCSI	Avg Write Exchange Completion Time (ms)	8.7050	6.9320	8.8120	8.6455
☑	5000000.0		G5-DP-2	Aggregate	SCSI	Max Write Size (bytes)	987,136.0000	987,136.0000	987,136.0000	987,136.0000
☑	50.0		G5-DP-2	Aggregate	SCSI	Pending Exchanges	3.0000	0.0000	5.0000	3.0833
☑	1.0		G5-DP-2	Aggregate	SCSI	Total Cmd to First Data Time (ms)	0.0000	0.0000	7.0710	0.1089
☑	2.0		G5-DP-2	Aggregate	SCSI	Avg Cmd to First Data Time (ms)	0.0000	0.0000	0.7070	0.0109
☑	10.0		G5-DP-2	Aggregate	SCSI	Max Cmd to First Data Time (ms)	0.0000	0.0000	2.2960	0.0292
☑	100.0		G5-DP-2	Aggregate	SCSI	Read Data Frames / Sec	0.0000	0.0000	40.0000	0.7692
☑	40.0		G5-DP-2	Aggregate	SCSI	Read Commands Issued	0.0000	0.0000	10.0000	0.1923
☑	60.0		G5-DP-2	Aggregate	SCSI	Read Commands Completed	0.0000	0.0000	10.0000	0.1923
☑	2.0		G5-DP-2	Aggregate	SCSI	Avg Read Exchange Completion Time (ms)	0.0000	0.0000	0.7310	0.0113
☑	10000.0		G5-DP-2	Aggregate	SCSI	Min Read Size (bytes)	0.0000	0.0000	4,096.0000	78.7692
☑	10000.0		G5-DP-2	Aggregate	SCSI	Max Read Size (bytes)	0.0000	0.0000	4,096.0000	78.7692
☑	10.0		G5-DP-2	Aggregate	SCSI	Min Read Exchange Completion Time (ms)	0.0000	0.0000	0.3650	0.0064

HD Capture

Channel	Loss of Sync Events	Loss of Signal Events	NOS and OLS Events	Extended Link Services Frames	Fibre Channel Service Frames	Fabric (SOFf) Frames	Basic Link Services Frames
Ch 1	0	0	0	0	0	0	0
Ch 2	0	0	0	0	0	0	0
Ch 1	0	0	0	0	0	0	0
Ch 2	0	0	0	0	0	0	0

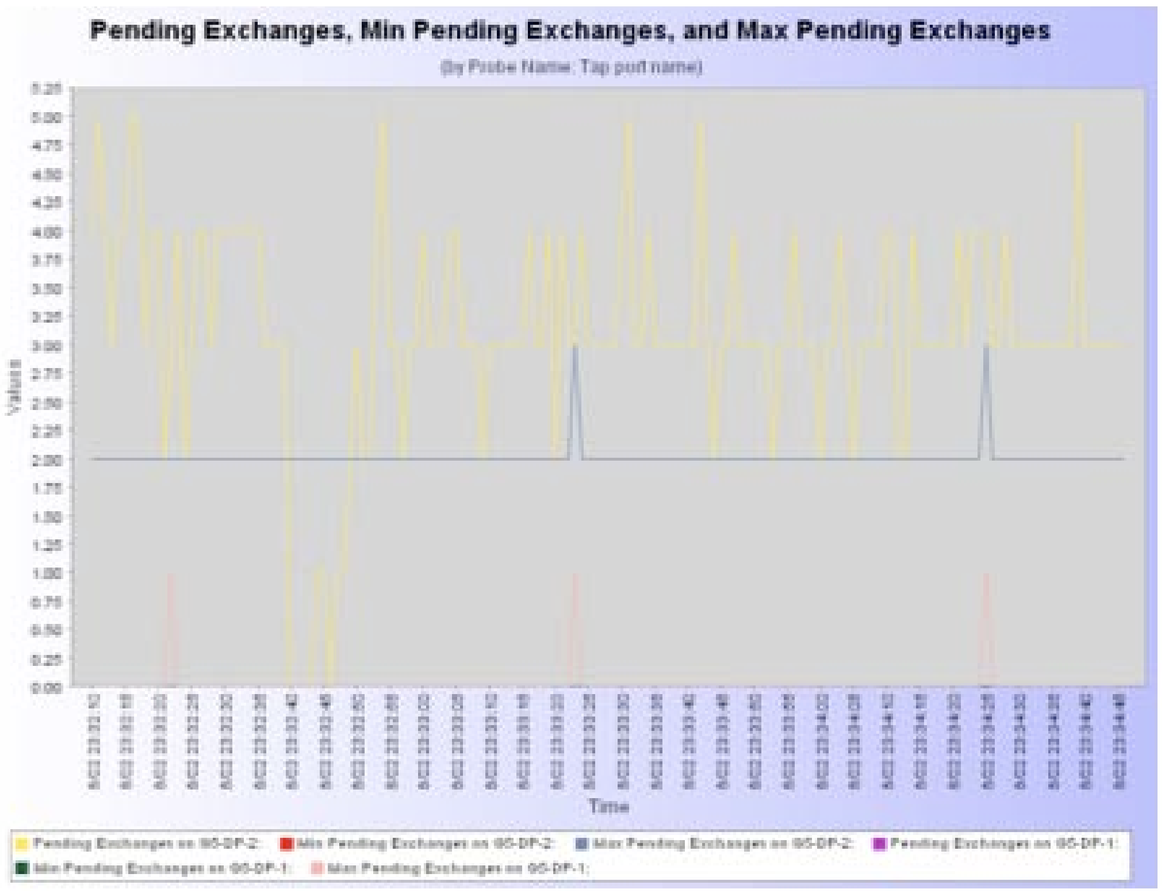
Channel	Link Control Frames	Link Up Events	#Check Condition Status Frames	#Other Bad Status Frames	Task Management Frames / Sec	Code Violations	Frame Errors
Ch 1	0	0	0	0	0	0	0
Ch 2	0	0	0	0	0	0	0
Ch 1	0	0	0	0	0	0	0
Ch 2	0	0	0	0	0	0	0

Channel	Logins	Logouts	Abort Sequence Frames	Notifications	Rejects	Busy	Accepts
Ch 1	0	0	0	0	0	0	0
Ch 2	0	0	0	0	0	0	0
Ch 1	0	0	0	0	0	0	0
Ch 2	0	0	0	0	0	0	0

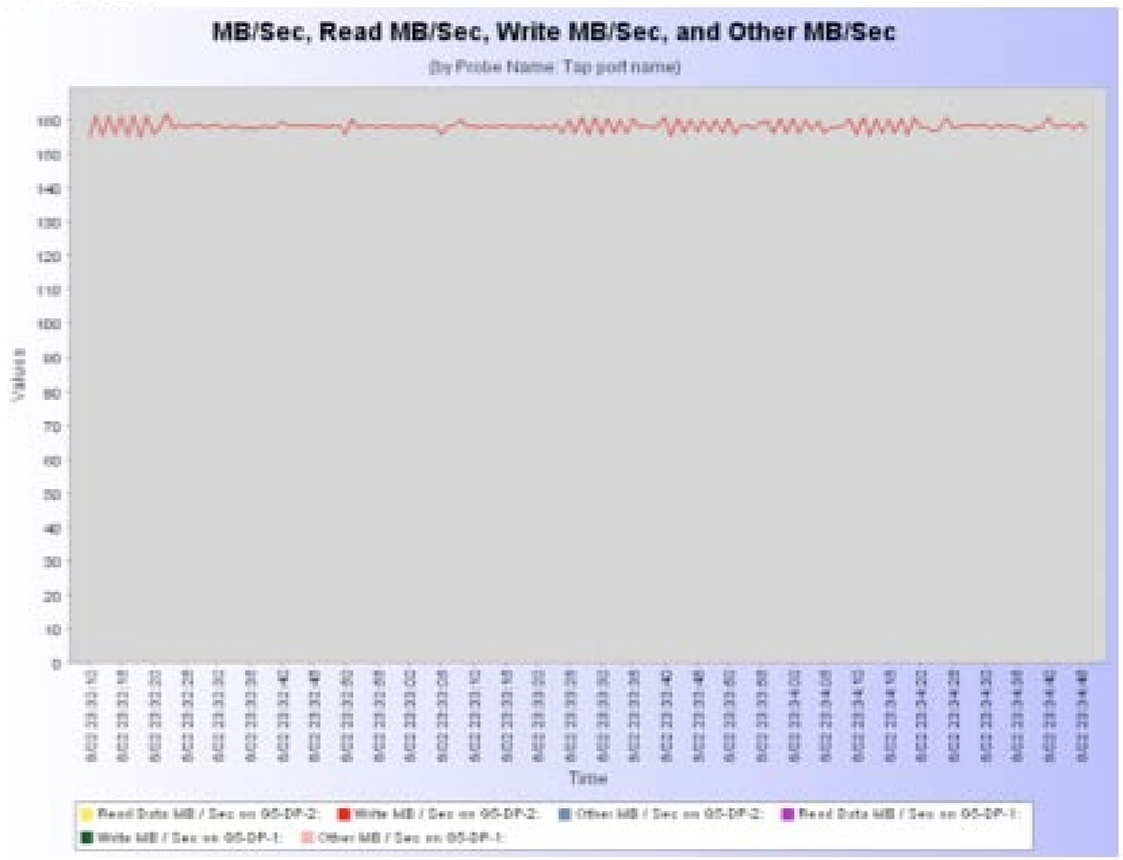


HD Capture

Pending Exchanges

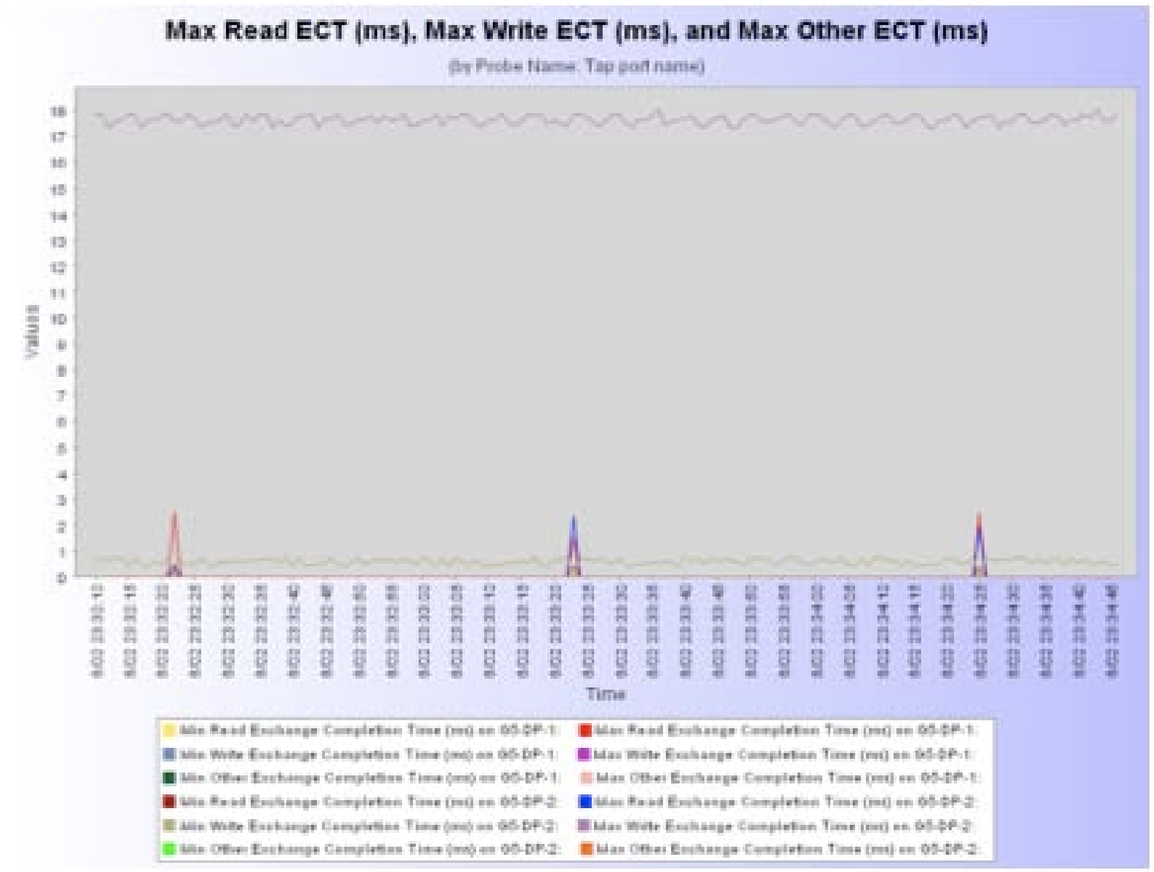


BandWidth

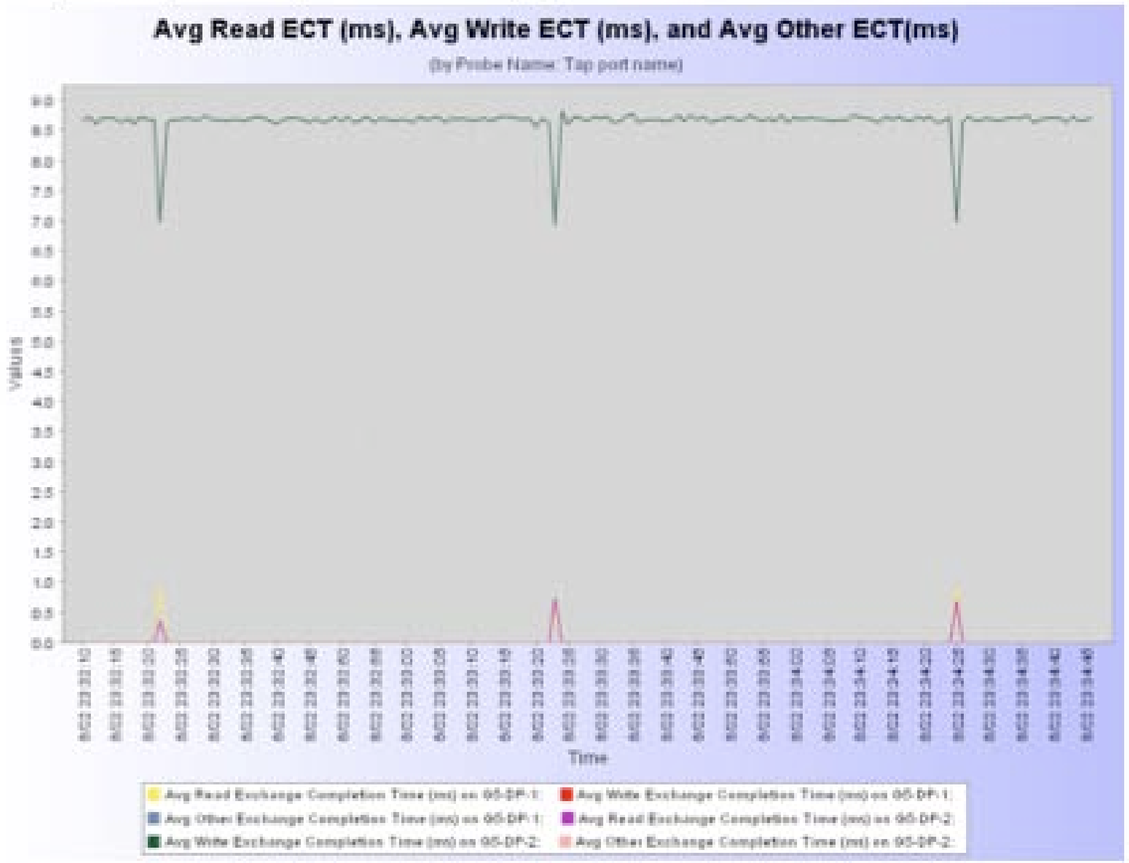


HD Capture

FCP (SCSI) Latency Continued

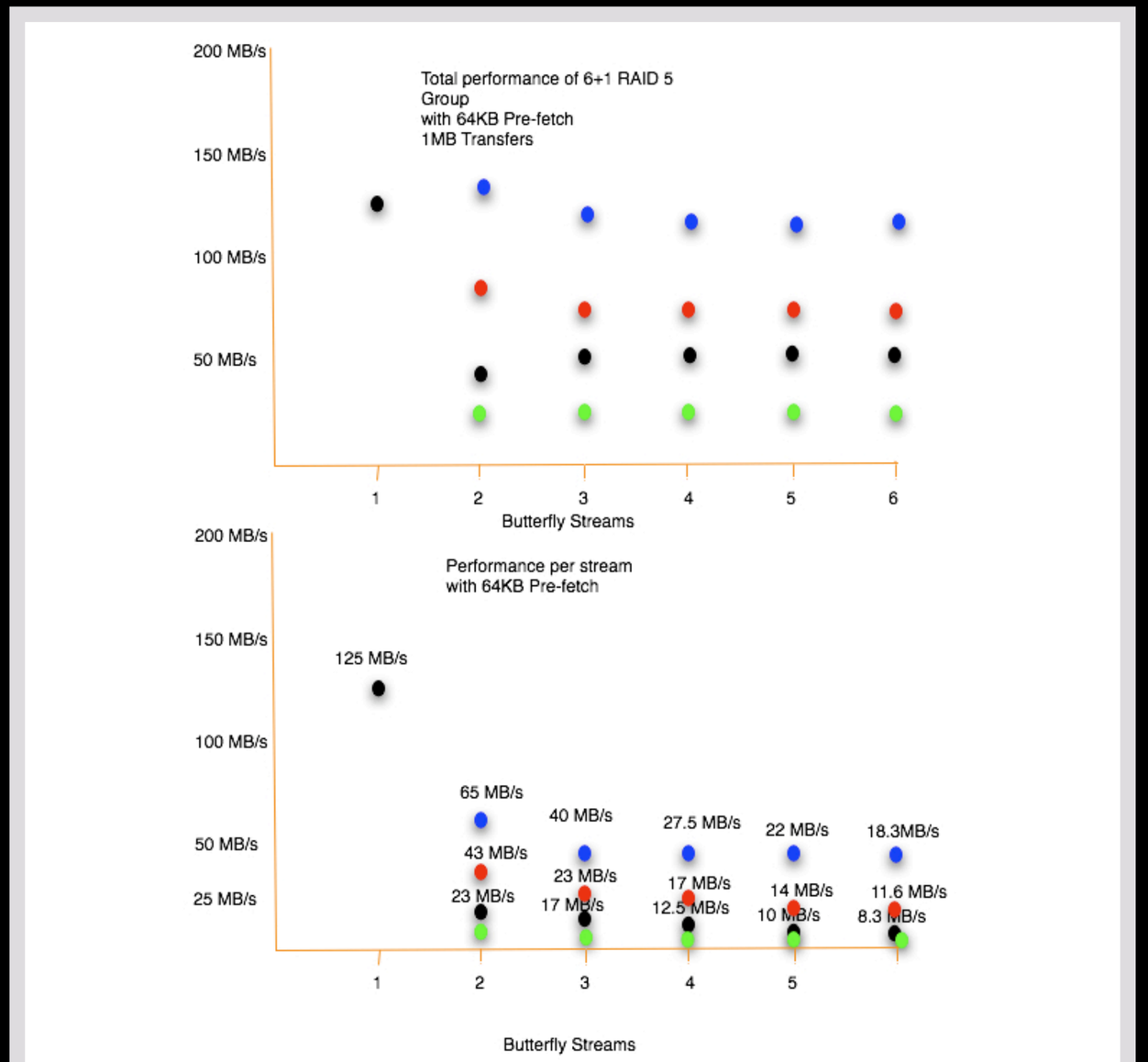


FCP (SCSI) Latency Continued (2)



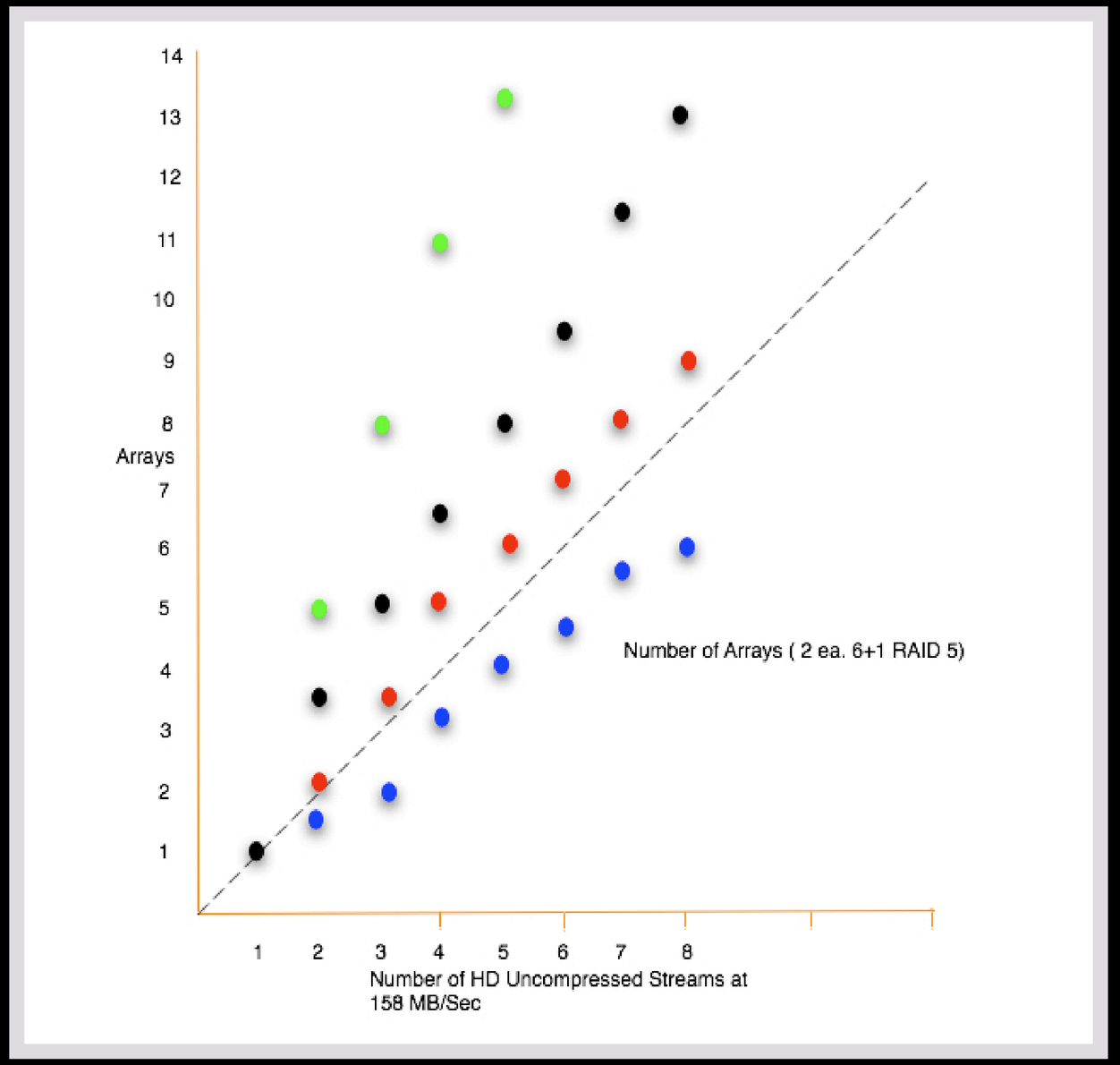
Test Results and Array Counts for HDTV

Method to determine number of arrays for # of streams



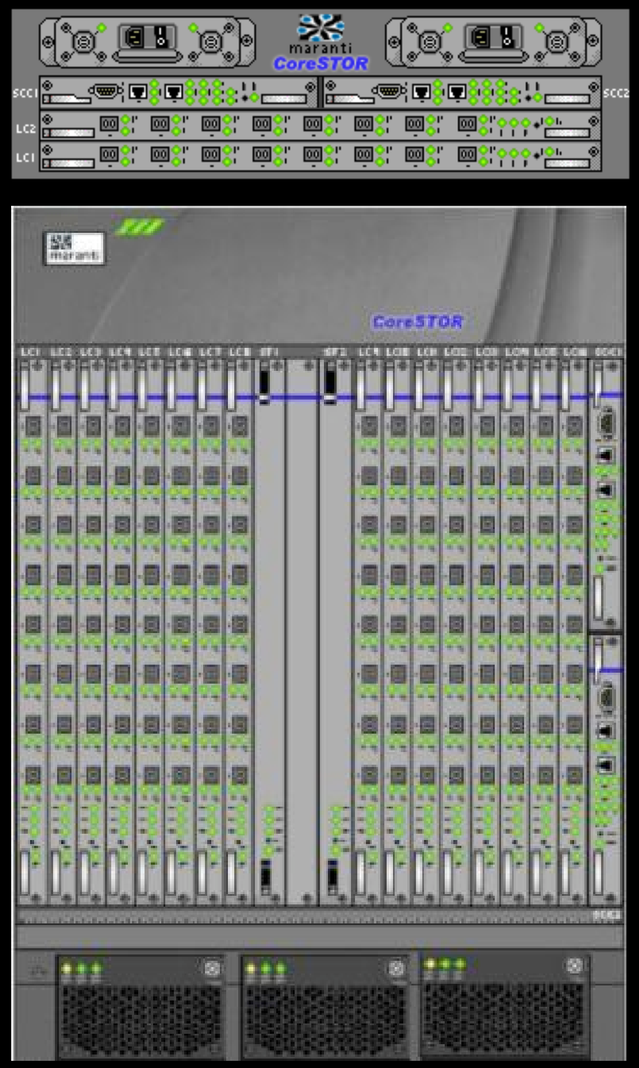
Test Results and Array Counts for HDTV

Method to determine number of arrays for # of streams



Large Scale Reference Design and Test Results

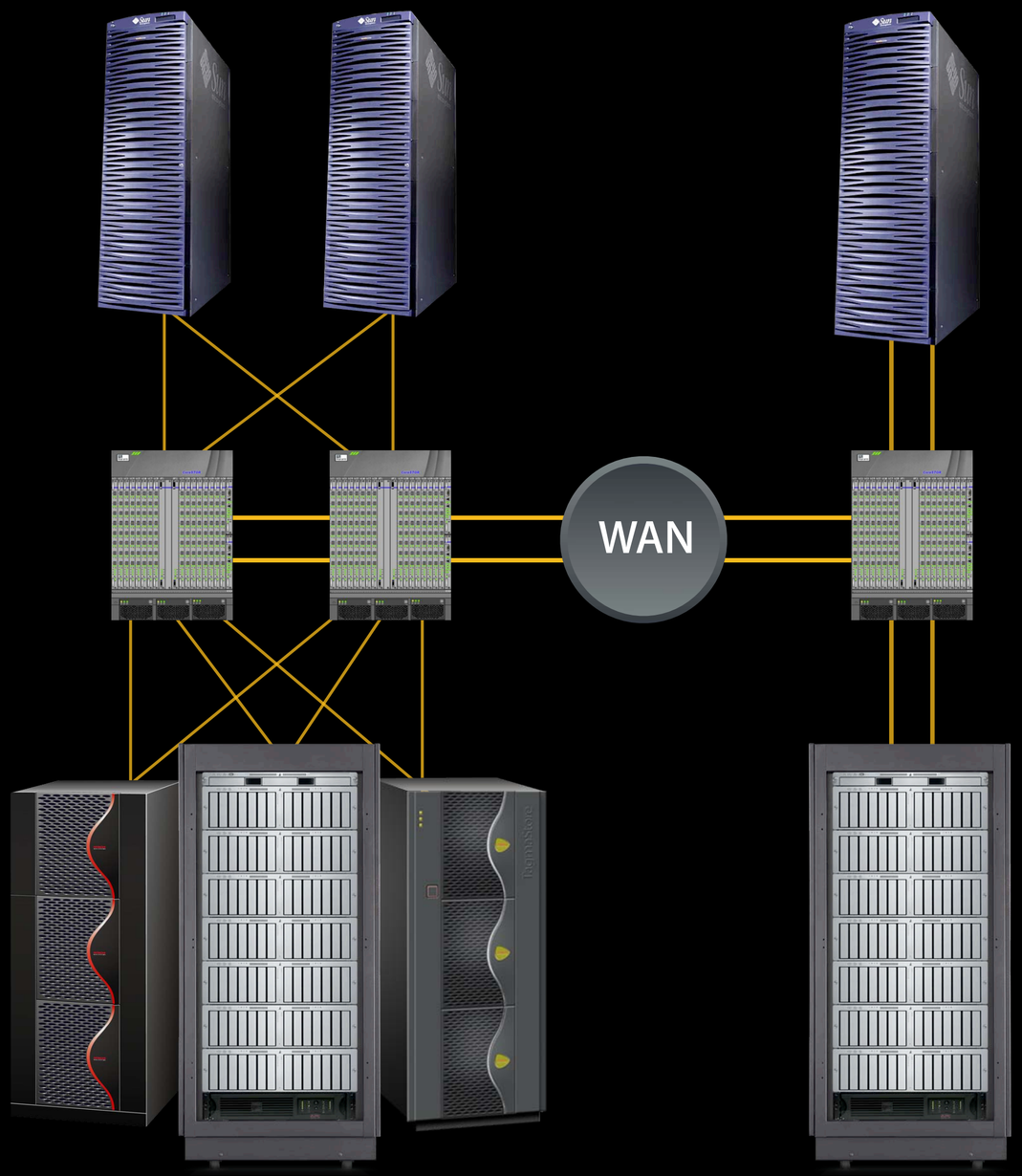
http://developer.apple.com/wwdc2005/



Large Scale Reference

Maranti CoreStor 2000 and 3000

- Network storage controller
- Non-Blocking 2Gb FC and iSCSI
- Adds
  - Storage virtualization
  - Mirroring
  - Replication
  - Snapshot
  - Single point of management



Large Scale Reference

Maranti CoreStor 2000 and 3000

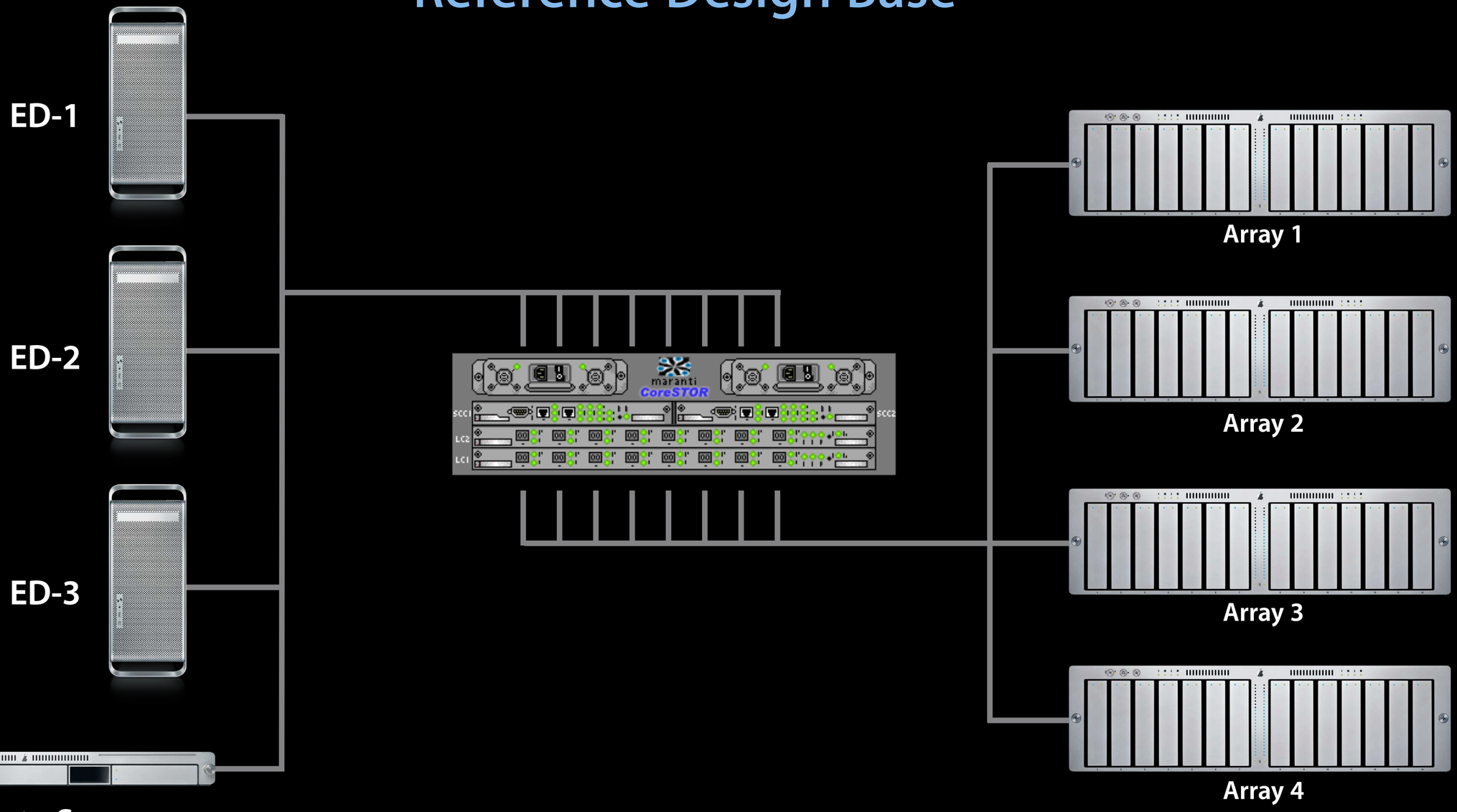
- Generally, great for
  - Tiered storage solutions with
    - EMC, HDS, IBM, Xserve RAID etc.
  - Local and remote data replication
  - Data mining
  - Quick data migration
  - Disk-to-disk backup

How About for Rich Media

- CoreSTOR Benefits
  - Increase performance
    - Trade off of capacity for performance
  - Workgroup isolation
    - Share media between groups without interaction
  - Redundancy
    - Mirror members give added data availability
  - Single management point

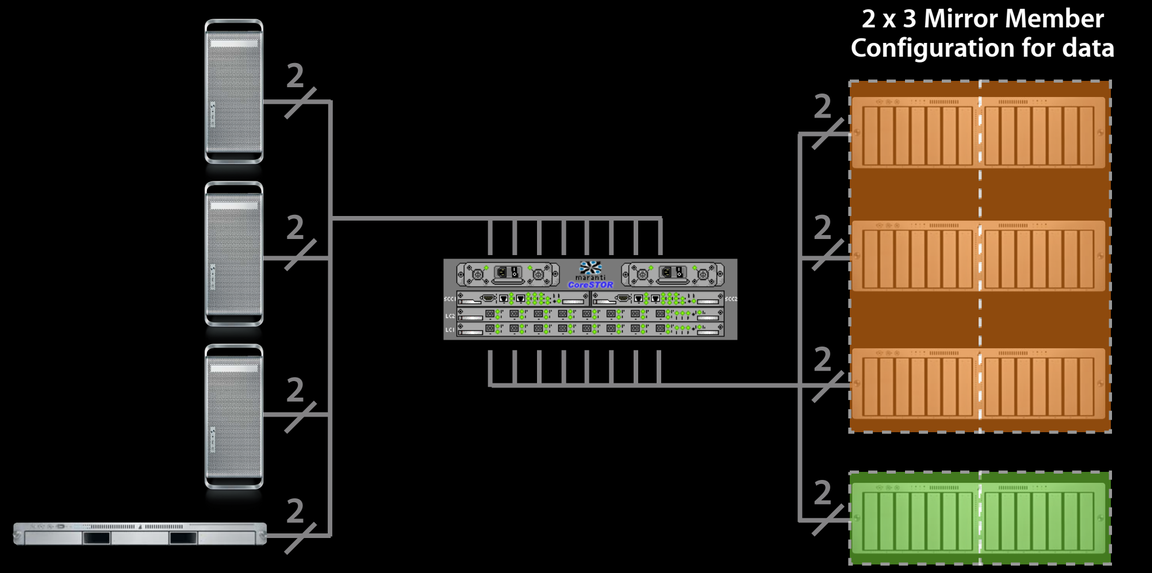
Large Scale Reference Design

Reference Design Base



Large Reference Design

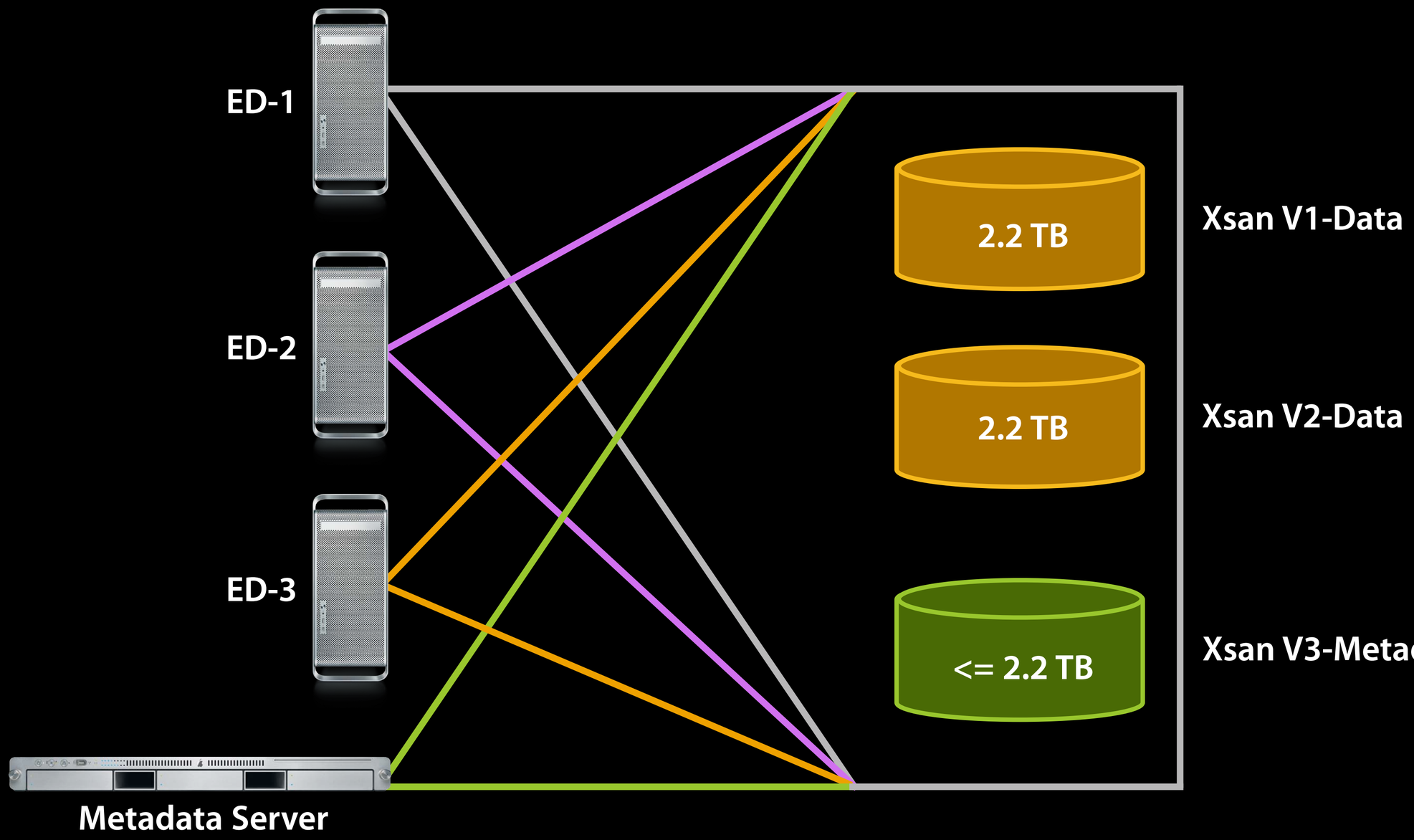
Mirror Member—2x3 Data and 1x2 Metadata



- Nomenclature
  - 2x3 means two volumes striped with 3 mirror members
  - 1x2 means one volume with 2 mirror members

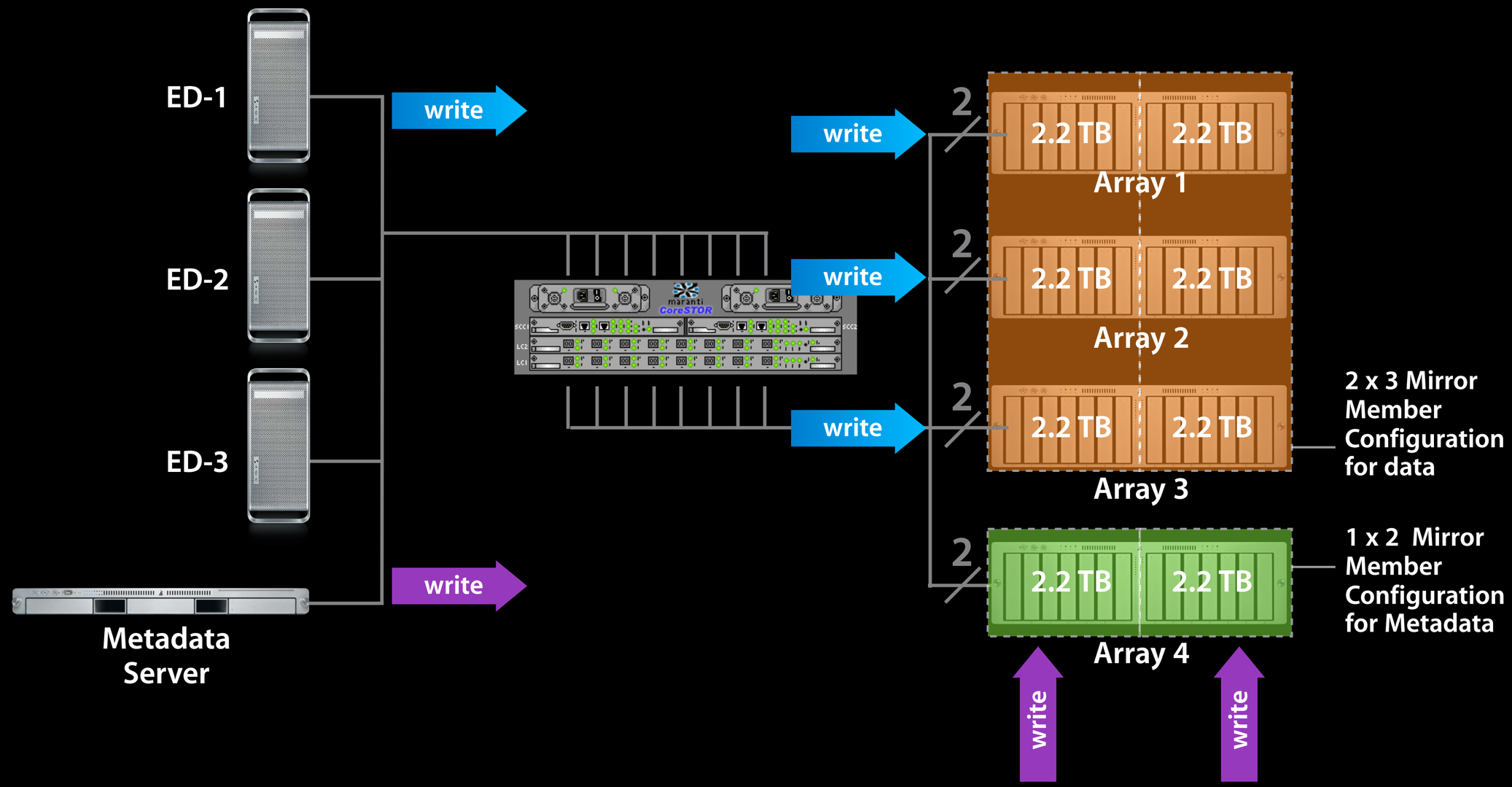
Large Reference Design

Host view of logical volumes



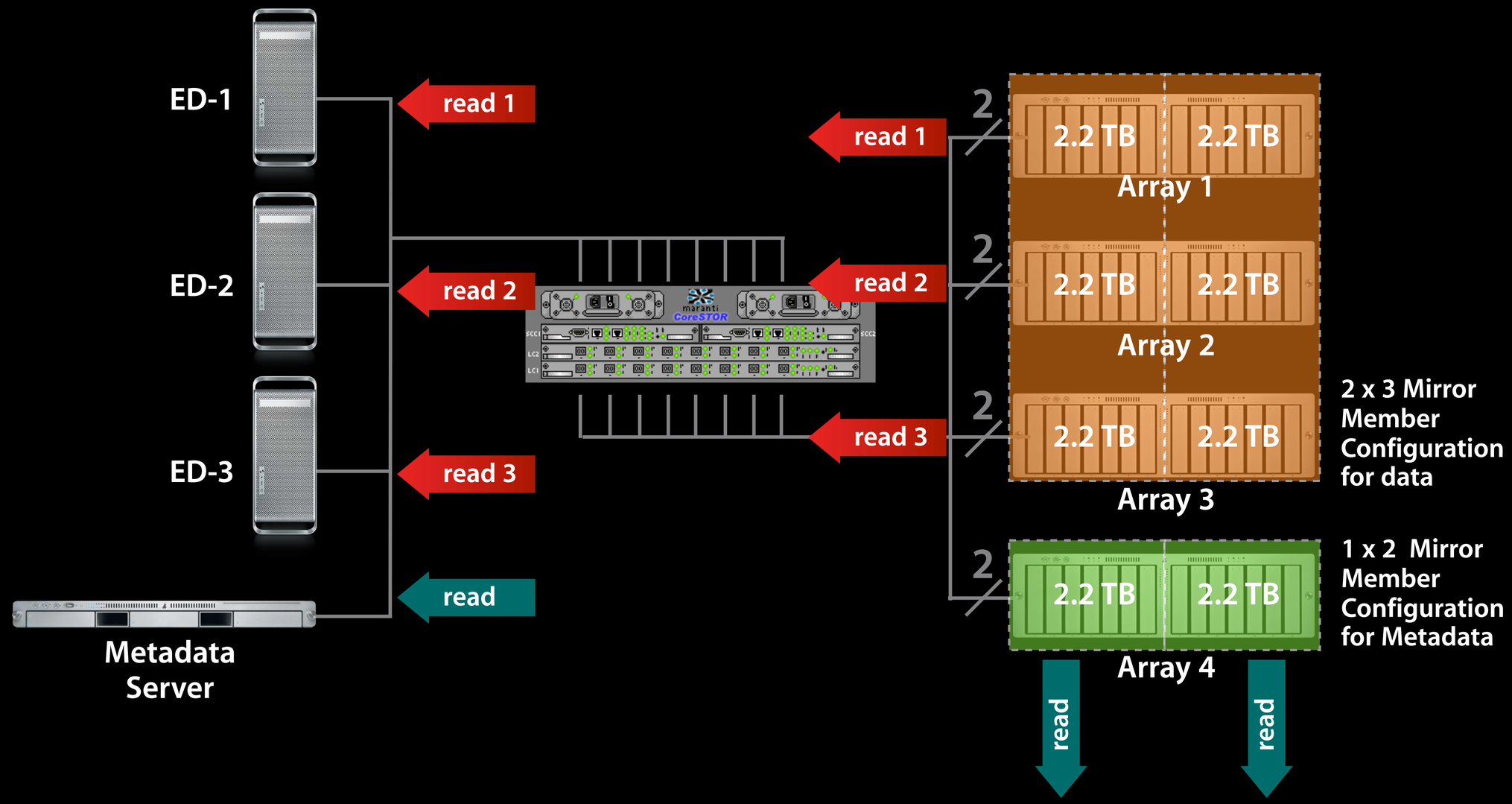
Large Reference Design

Multi-casting writes to mirror members



Large Reference Design

Increased read performance and added redundancy

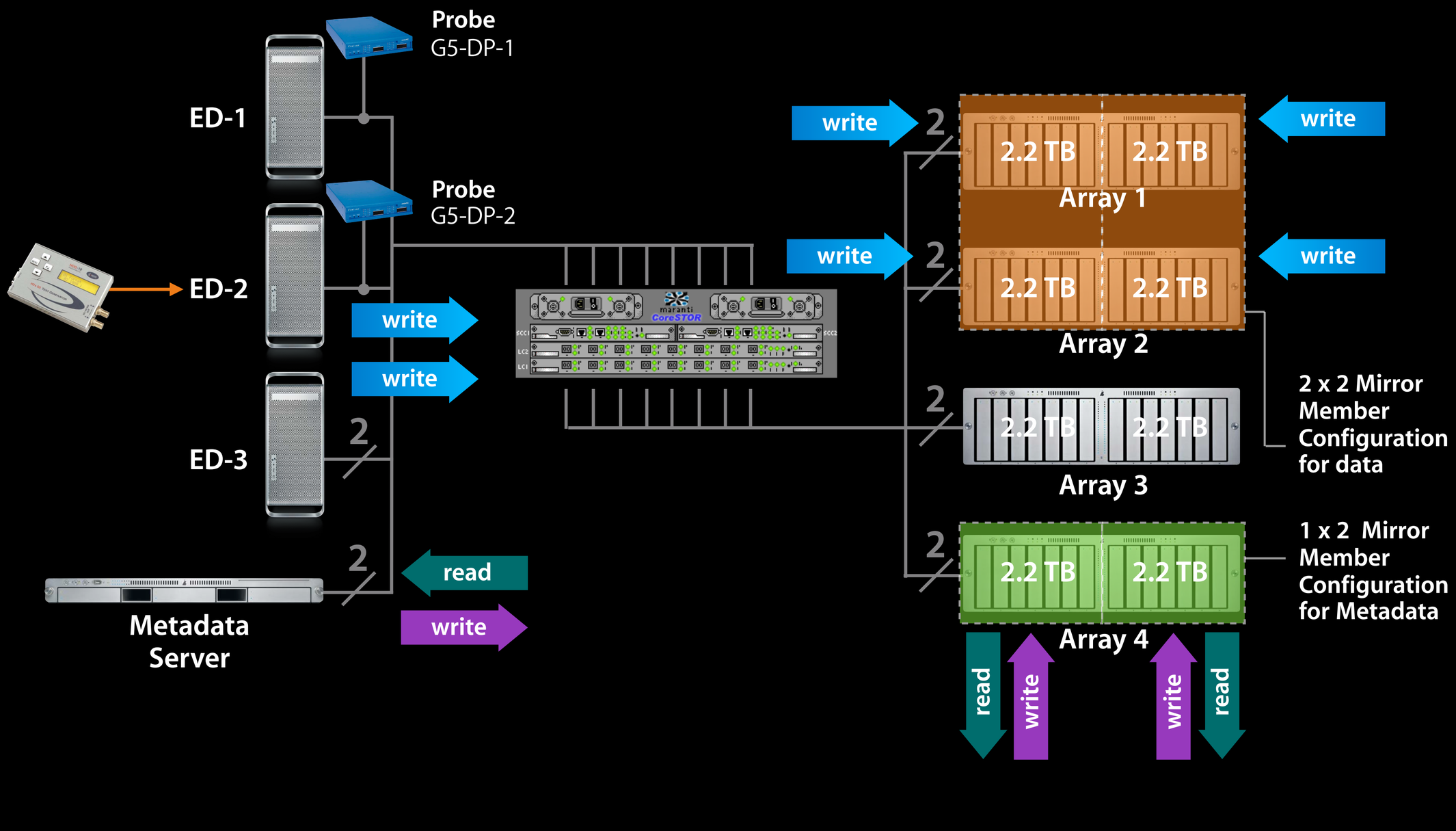


Next, Test with HDTV

http://developer.apple.com/wwdc2005/

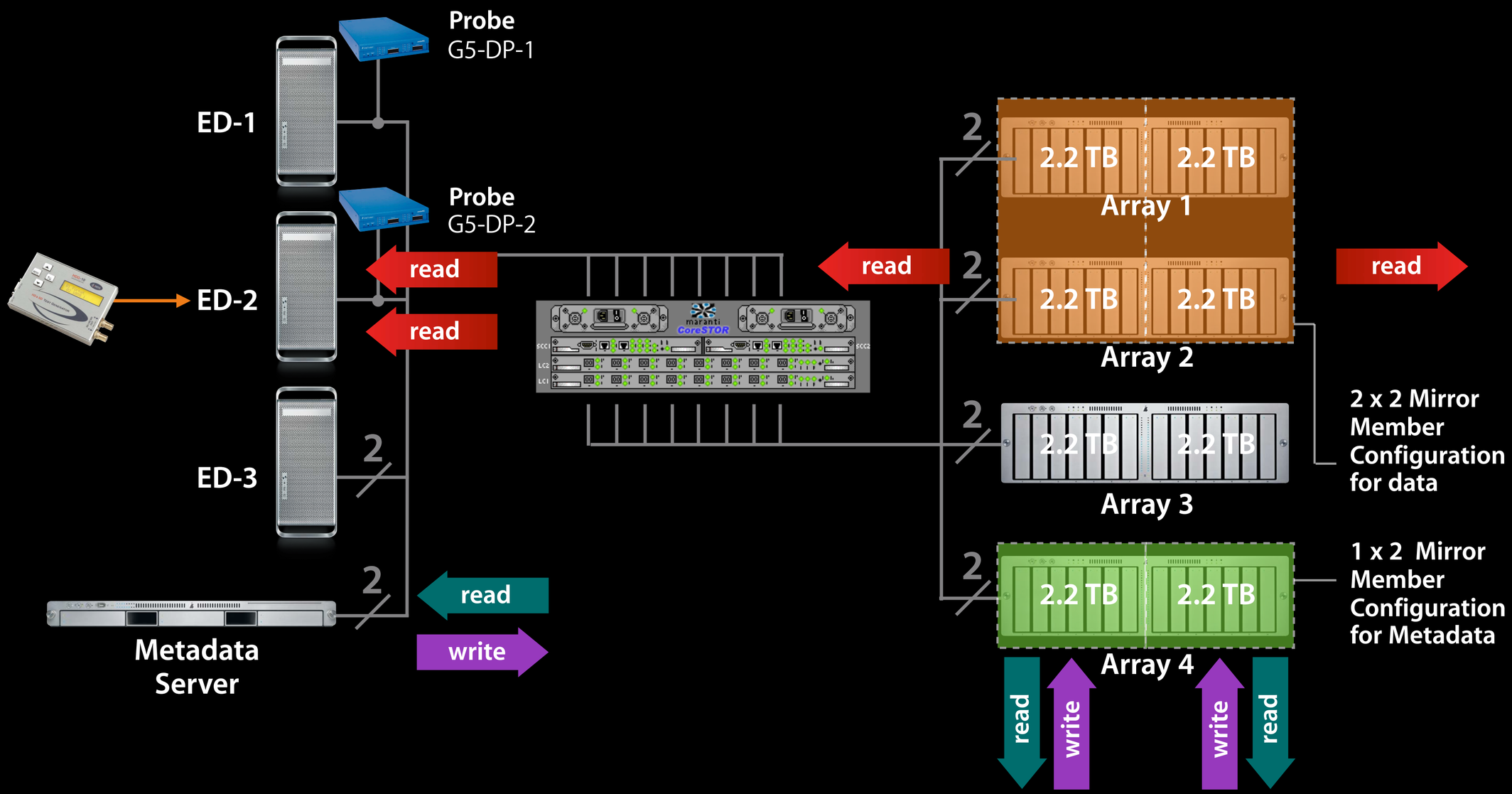
Test Results

Capture HDTV on two Arrays



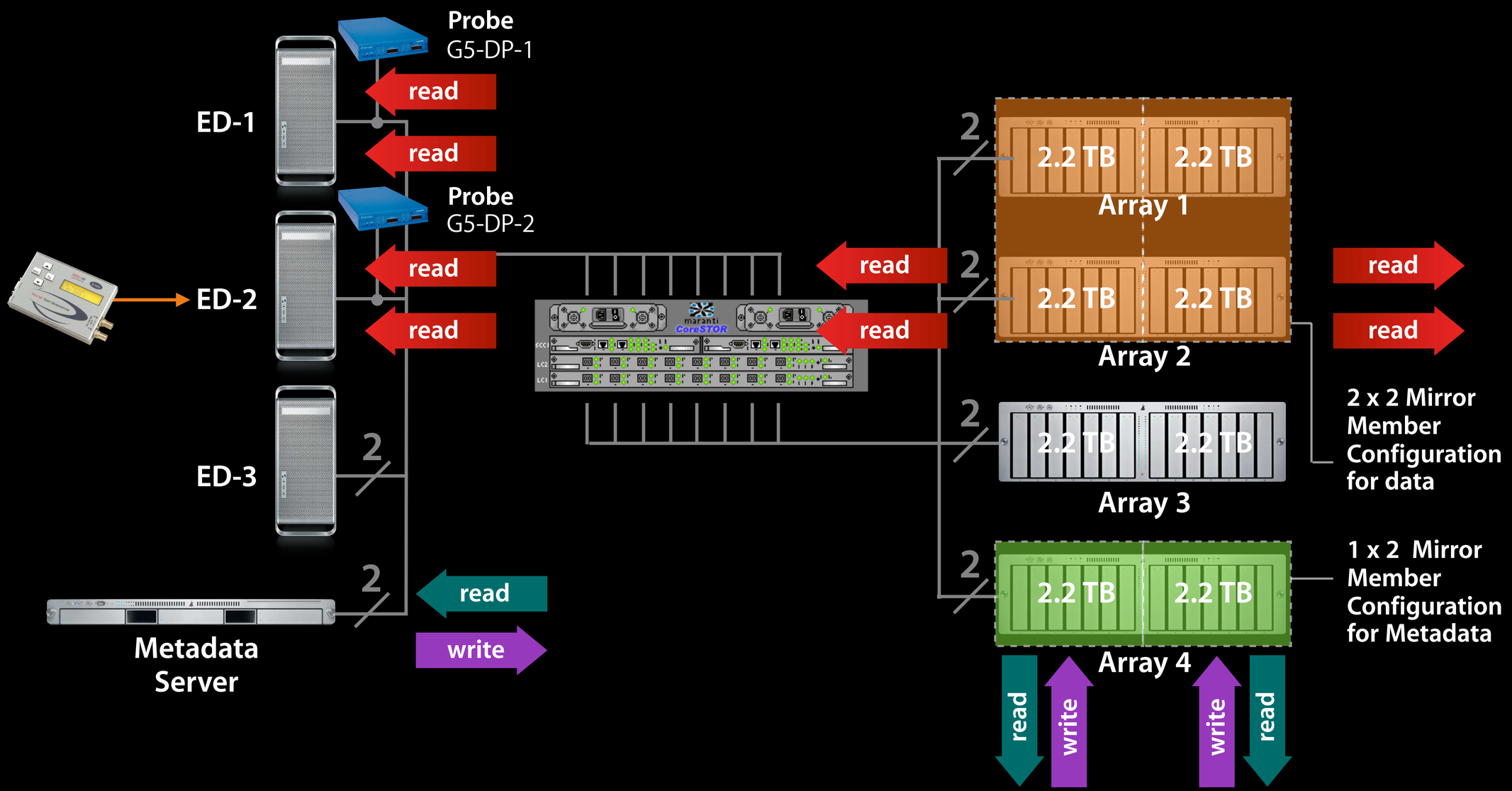
Test Results

Read back HDTV real-time from one member



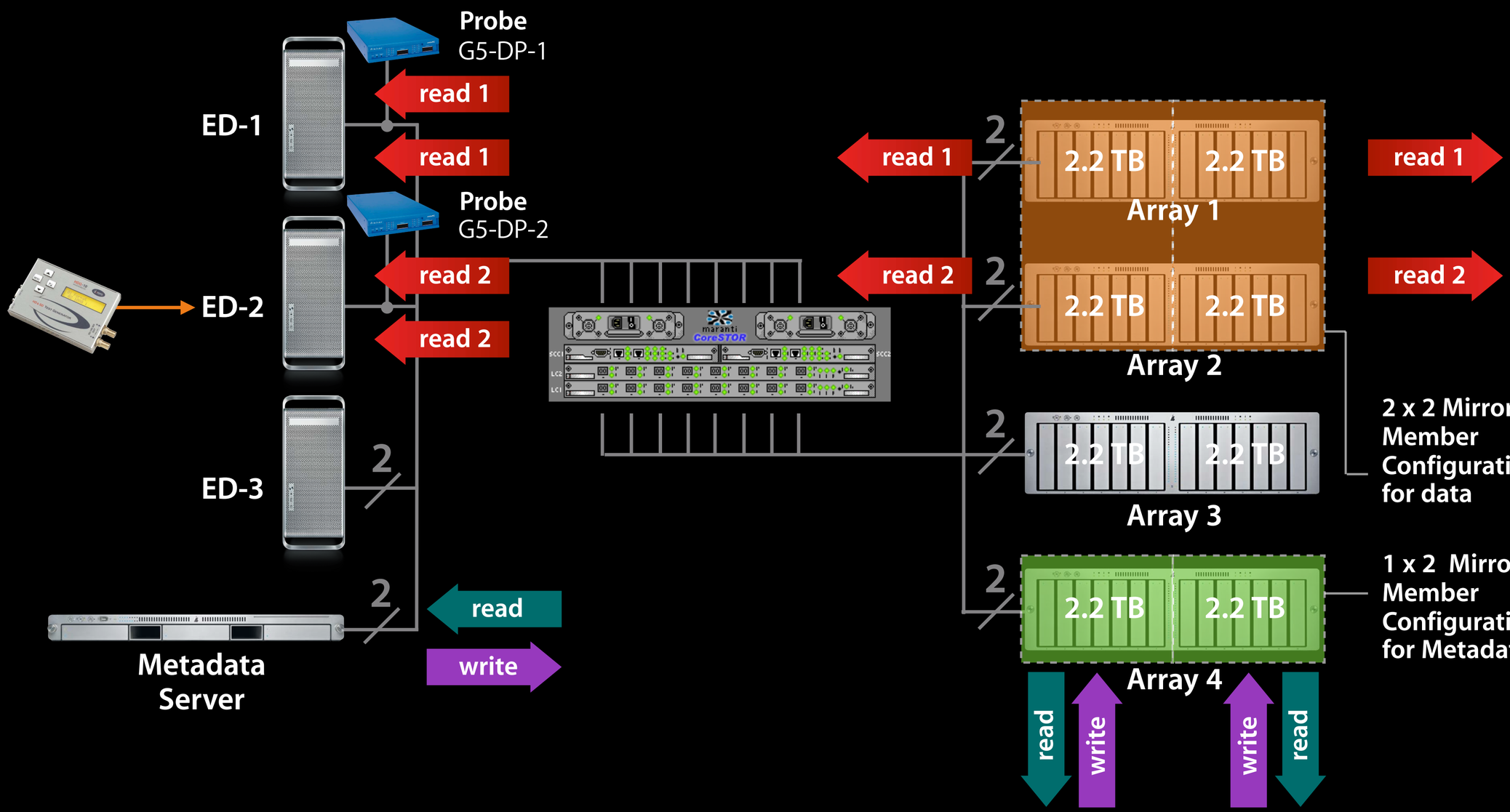
Test Results

Attempt to read dual streams from one Array



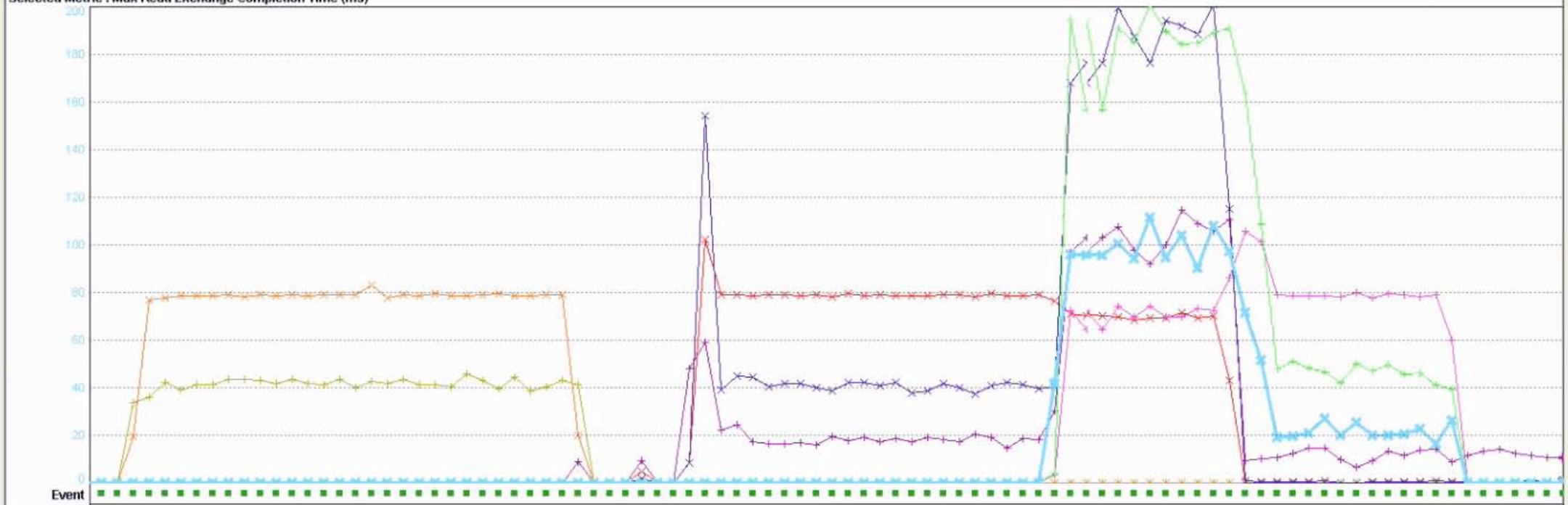
Test Results

Change CoreStor configuration, preferred read mode



06/02/05 16:37:03 PDT | 06/02/05 16:38:36 | 06/02/05 16:39:40 PDT

Selected Metric : Max Read Exchange Completion Time (ms)



No Event | Event from Plotted Probes | Event from Unplotted Probes

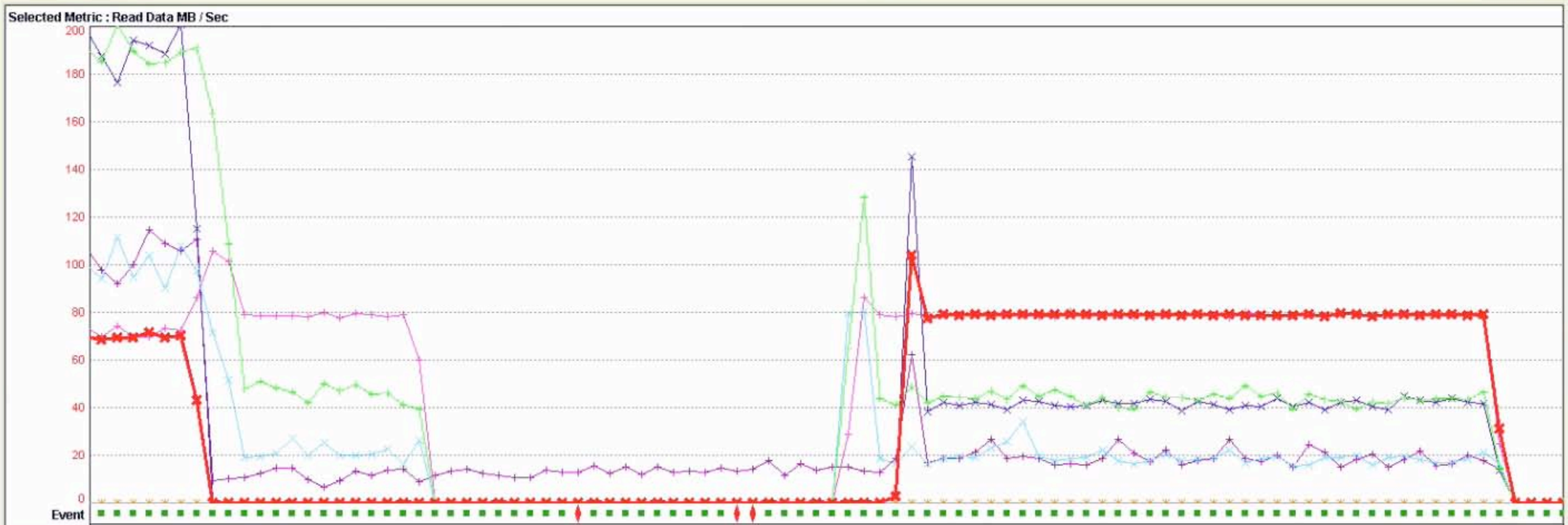
Graph Legend | FC/SCSI version 2.2/FcCh - Event Counters | Metric Selection

final take config finisar nw — New

Display	Scale	Color	Probe	Aggregation	Metric Set	Counter	Current	Minimum	Maximum	Average
☑	200.0		G5-DP-1	Aggregate	SCSI	Write MB / Sec	0.0000	0.0000	0.0000	0.0000
☑	200.0		G5-DP-2	Aggregate	SCSI	Write MB / Sec	0.0000	0.0000	83.2227	23.1418
☑	60.0		G5-DP-2	Aggregate	SCSI	Max Write Exchange Completion Time (ms)	0.0000	0.0000	13.8470	3.8622
☑	200.0		G5-DP-2	Aggregate	SCSI	Read Data MB / Sec	0.0781	0.0000	102.3330	26.9324
☑	200.0		G5-DP-2	Aggregate	SCSI	Max Read Exchange Completion Time (ms)	13.7890	0.0000	114.8470	19.6669
☑	10000.0		G5-DP-2	Aggregate	SCSI	Total Read Exchange Completion Time (ms)	37.7350	0.0000	10,048.4610	1,534.7821
☑	200.0		G5-DP-1	Aggregate	SCSI	Read Data MB / Sec	0.0000	0.0000	106.1602	19.9126
☑	200.0		G5-DP-1	Aggregate	SCSI	Max Read Exchange Completion Time (ms)	0.0000	0.0000	111.7900	15.1179
☑	10000.0		G5-DP-1	Aggregate	SCSI	Total Read Exchange Completion Time (ms)	0.0000	0.0000	10,055.7340	1,439.7607

Print... | Set Bookmark | Reset graph

06/02/05 16:37:03 PDT | 06/02/05 16:39:40 | 06/02/05 16:39:40 PDT



Graph Legend | FC/SCSI version 2.2/FcCh - Event Counters | Metric Selection

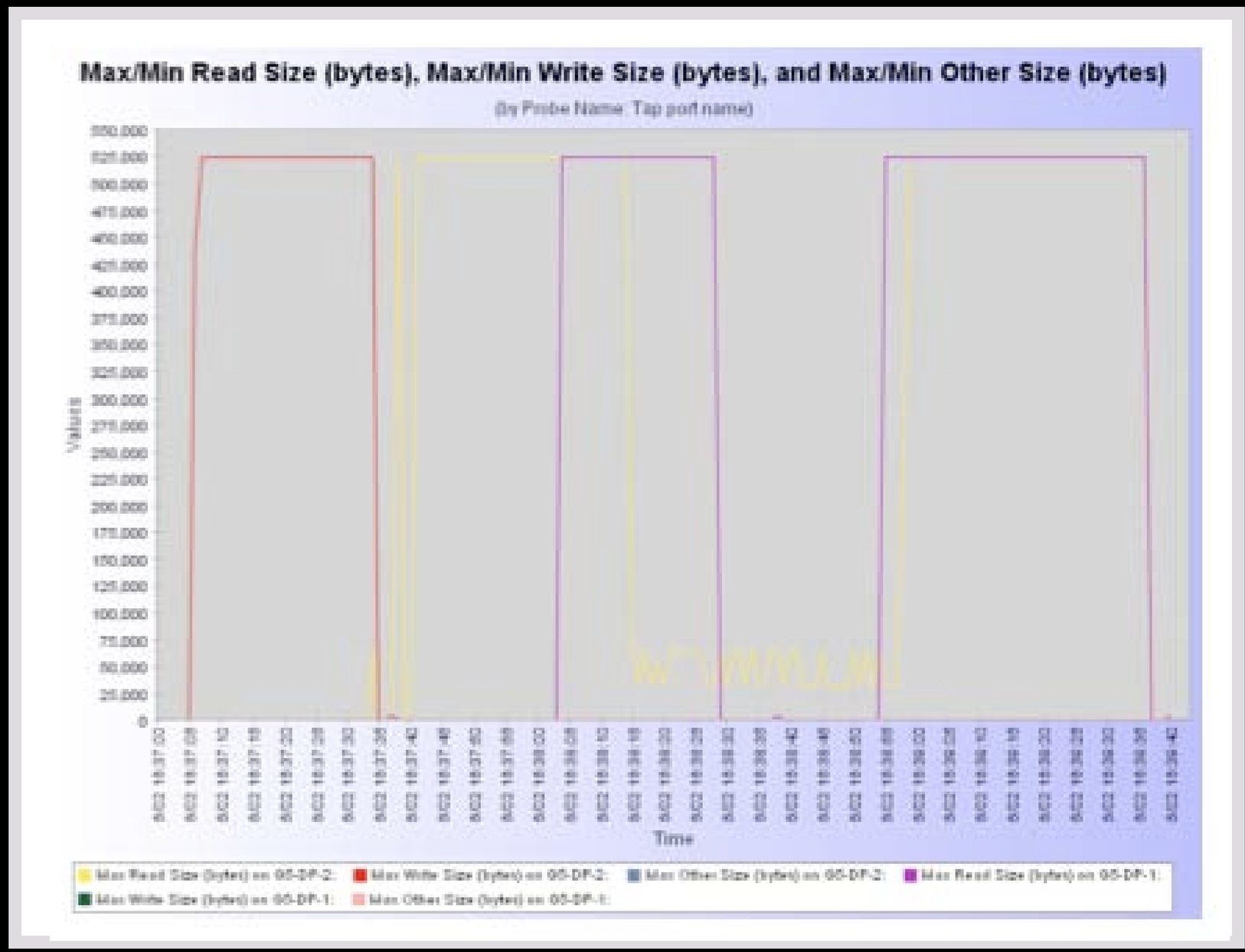
final take config finisar nw

New

Display	Scale	Color	Probe	Aggregation	Metric Set	Counter	Current	Minimum	Maximum	Average
☑	200.0		G5-DP-1	Aggregate	SCSI	Write MB / Sec	0.0000	0.0000	0.0000	0.0000
☑	200.0		G5-DP-2	Aggregate	SCSI	Write MB / Sec	0.0000	0.0000	83.2227	13.7679
☑	60.0		G5-DP-2	Aggregate	SCSI	Max Write Exchange Completion Time (ms)	0.0000	0.0000	13.8470	2.2978
☑	200.0		G5-DP-2	Aggregate	SCSI	Read Data MB / Sec	0.0000	0.0000	104.2578	34.9413
☑	200.0		G5-DP-2	Aggregate	SCSI	Max Read Exchange Completion Time (ms)	0.0000	0.0000	114.8470	18.5955
☑	10000.0		G5-DP-2	Aggregate	SCSI	Total Read Exchange Completion Time (ms)	0.0000	0.0000	10,048.4610	1,446.2932
☑	200.0		G5-DP-1	Aggregate	SCSI	Read Data MB / Sec	0.0234	0.0000	106.1602	32.2646
☑	200.0		G5-DP-1	Aggregate	SCSI	Max Read Exchange Completion Time (ms)	0.3380	0.0000	111.7900	14.9469
☑	10000.0		G5-DP-1	Aggregate	SCSI	Total Read Exchange Completion Time (ms)	1.8790	0.0000	10,055.7340	1,470.2682

Print... | Set Bookmark | Reset graph

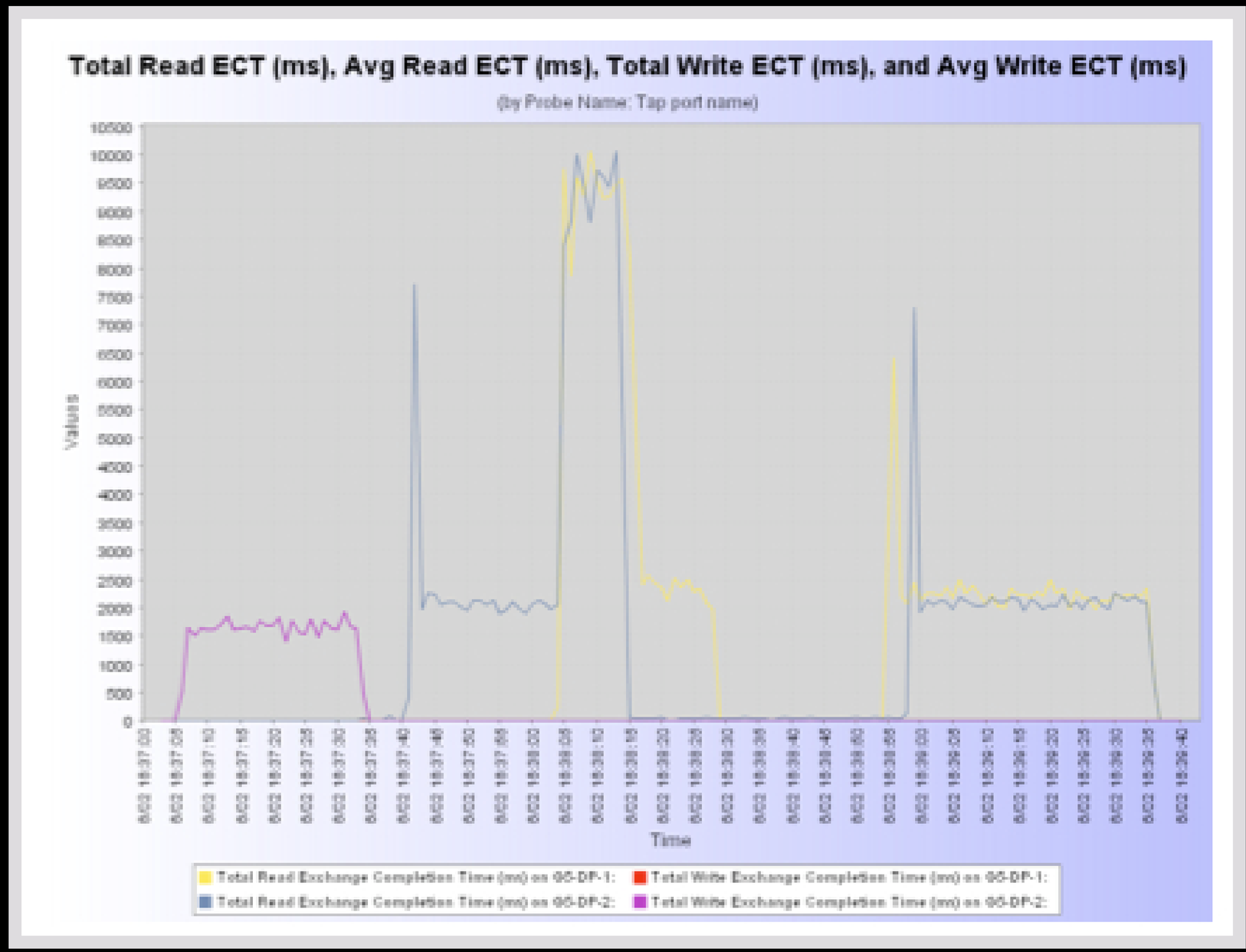
I/O Size


Max/Min Read Size (bytes), Max/Min Write Size (bytes), and Max/Min Other Size (bytes)
(by Probe Name: Tap port name)
Values
Time
Max Read Size (bytes) on 05-DP-2:
Max Write Size (bytes) on 05-DP-2:
Max Other Size (bytes) on 05-DP-2:
Max Read Size (bytes) on 05-DP-1:
Max Write Size (bytes) on 05-DP-1:
Max Other Size (bytes) on 05-DP-1:

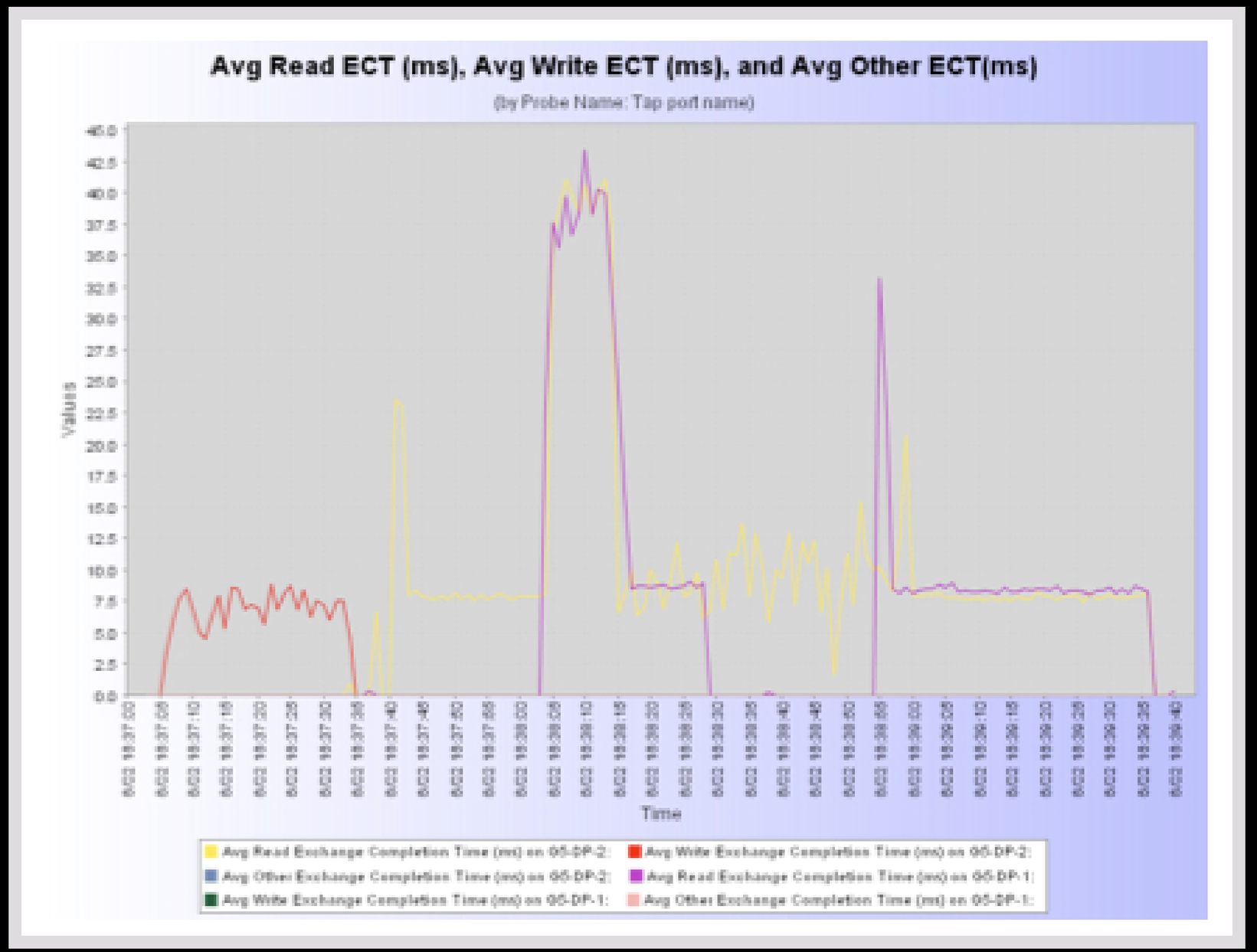
Bandwidth

MB/Sec, Read MB/Sec, Write MB/Sec, and Other MB/Sec
(by Probe Name: Tap port name)
Values
Time
Read Data MB / Sec on 05-DP-2:
Write MB / Sec on 05-DP-2:
Other MB / Sec on 05-DP-2:
Read Data MB / Sec on 05-DP-1:
Write MB / Sec on 05-DP-1:
Other MB / Sec on 05-DP-1:

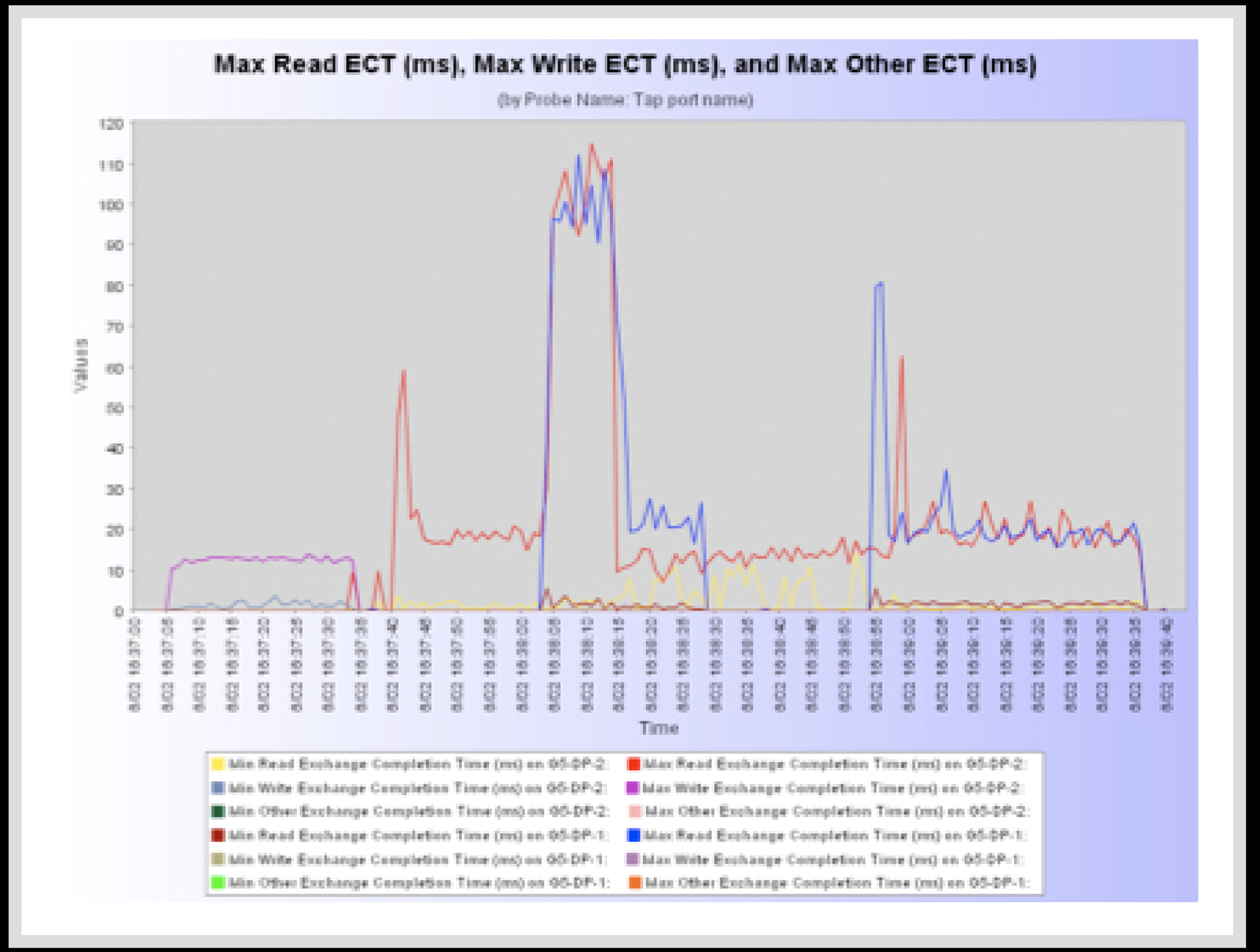
FCP (SCSI) Latency



FCP (SCSI) Latency Continued (2)

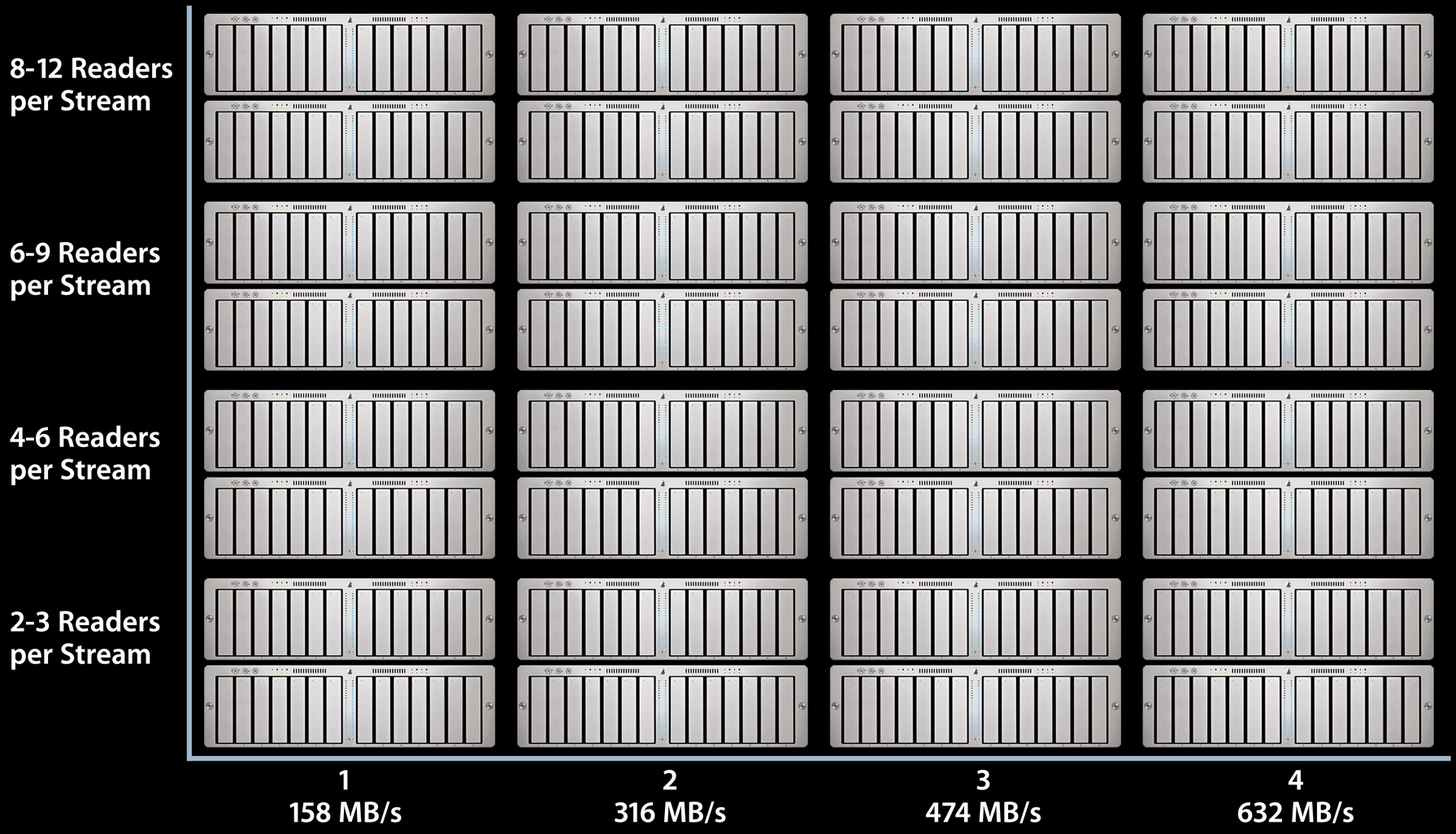


6. FCP (SCSI) Latency Continued

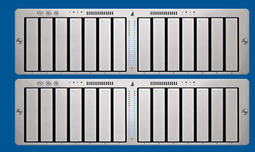


Test Results

Scaling Xserve RAIDS for multi-stream HDTV

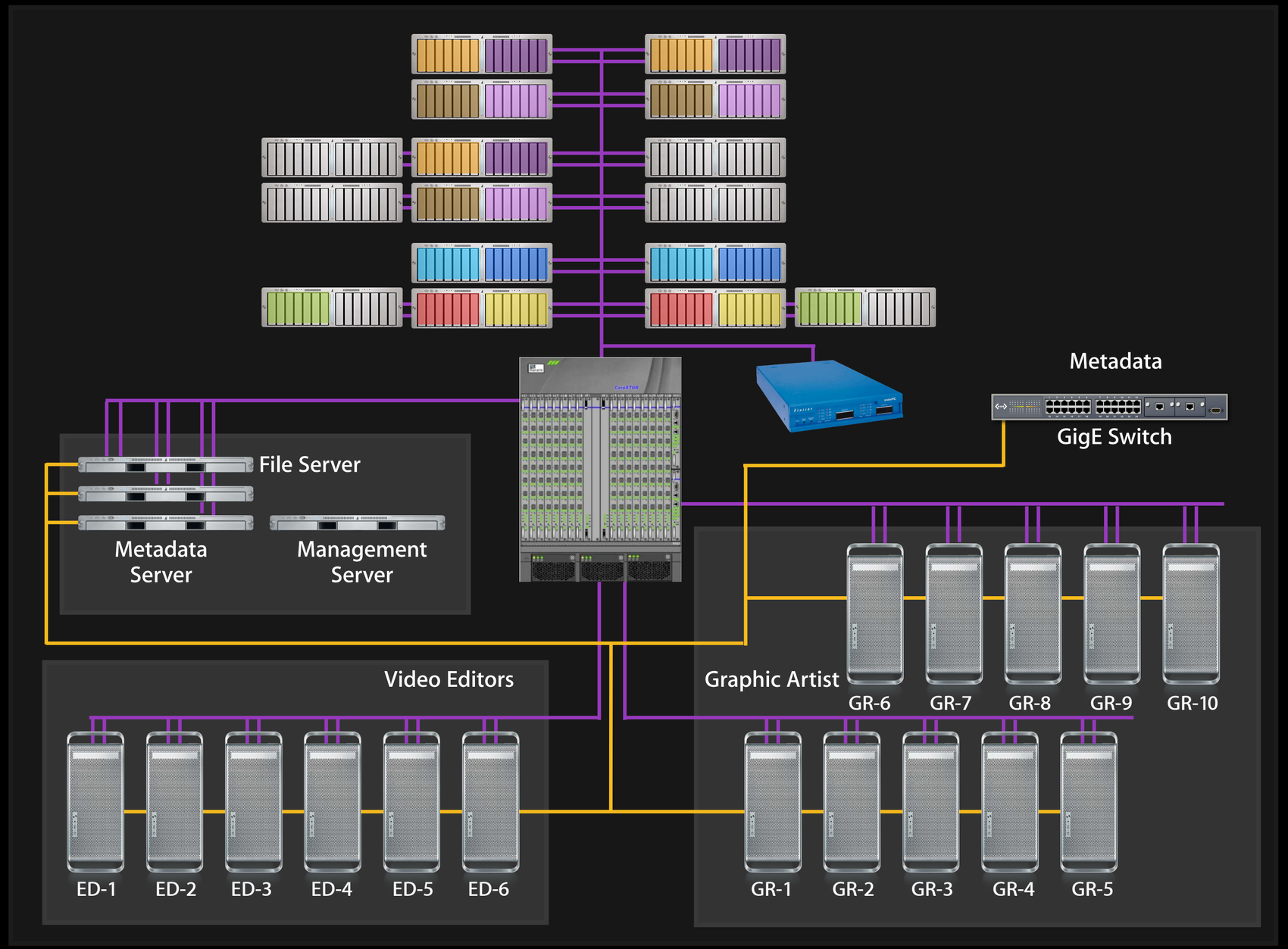


One Array = 1 Xsan writer or 1 reader



Two Arrays = 1 Xsan writer and 1 reader, or 2 readers with some writes 3 readers with no writes

Reference Implementation



In Conclusion…

- Deployment starts with a good understanding of customer needs
- Leverage off reference designs
  - Xserve RAID a great building block for speed and predictability
- Know your products well by using test and monitoring tools
  - Your success to scale depends this knowledge
- Build a small system to test applications
- Add concurrent access
- Scale to large deployment
- Test and monitor
- Add pro-active management

Contact Information

- For Full SAN Solution test result bundle or any questions or comments
  - Please email:
    - wwdc@sansolutions.com

- Thank you, The SAN solutions team

- Next Questions and answers panel
  - Richard Jew, Finisar Corporation
  - Gary Archer, Maranti Networks, Inc.
  - Drew Hall, SAN Solutions, Inc.
  - Harry Aine, SAN Solutions, Inc.

Q&A

http://developer.apple.com/wwdc2005/

WWDC2005

Worldwide Developers Conference